Metric-based Data Quality Assessment -

Developing and Evaluating a Probability-based Currency Metric

Authors:

Heinrich, Bernd, Department of Management Information Systems, University of Regensburg, Universitätsstraße 31, D-93040 Regensburg, Germany, bernd.heinrich@ur.de

Mathias, Klier, Department of Management Information Systems, University of Regensburg, Universitätsstraße 31, D-93040 Regensburg, Germany, mathias.klier@ur.de

Citation: Bernd Heinrich, Mathias Klier, Metric-based data quality assessment — Developing and evaluating a probability-based currency metric, Decision Support Systems, Volume 72, April 2015, Pages 82-96, ISSN 0167-9236, http://dx.doi.org/10.1016/j.dss.2015.02.009.

(http://www.sciencedirect.com/science/article/pii/S0167923615000299)

Metric-based Data Quality Assessment -

Developing and Evaluating a Probability-based Currency Metric

Abstract:

Data quality assessment has been discussed intensively in the literature and is critical in business. The importance of using up-to-date data in business, innovation, and decision-making processes has revealed the need for adequate metrics to assess the currency of data in information systems. In this paper, we propose a data quality metric for currency that is based on probability theory. Our metric allows for a reproducible configuration and a high level of automation when assessing the currency of attribute values. The metric values represent probabilities and can be integrated into a decision calculus (e.g., based on decision theory) to support decision-making. The evaluation of our metric consists of two main steps: (1) we define an instantiation of the metric for a real-use situation of a German mobile services provider to demonstrate both the applicability and the practical benefit of the approach; (2) we use publicly available real world data provided by the Federal Statistical Office of Germany and the German Institute of Economic Research to demonstrate its feasibility by defining an instantiation of the metric and to evaluate its strength (compared to existing approaches).

Keywords: Data quality, Data quality assessment, Data quality metric, Currency of data

1. INTRODUCTION

Data quality issues are discussed intensively in the literature and are critical in business. High-quality data in information systems (IS) are needed as a basis for business, innovation, and decision-making processes (Al-Hakim 2007, Ofner et al. 2012). Thus, poor data quality often results in bad decisions and economic losses (Even et al. 2010, Forbes 2010, Gelman 2010, Qas 2013). In addition, great effort is required to ease or solve data quality problems (Even and Shankaranarayanan 2007, IBM 2012). The growing relevance of data quality has also revealed the need for adequate assessment (e.g., IBM 2012, Qas 2013). Quantifying data quality (e.g., quality of customer data) is essential for taking into account data quality aspects in decision-making (e.g., to select the customers to be addressed in a mailing campaign considering the quality of the address data stored). Moreover, assessing data quality constitutes an indispensable step toward the ability to decide whether a data quality measure (e.g., address data cleansing) should be taken from an economic perspective. In this context, it is necessary to quantify and consider the effects of measures with respect to the data quality level – a fact that is

often illustrated as part of the data quality loop (for details cf. e.g., Heinrich et al. 2009).

A recent report (cf. Qas 2013) has revealed that one of the most common data defects is outdated data, which primarily results in wasted budgets, loss of potential customers, and reduced customer satisfaction. For example, two-thirds of surveyed organizations observe several problems in the context of customer relationship management, such as sending mailings to the wrong address or sending the same mailing to the customer multiple times; this indicates that outdated address and customer data negatively affect customer perceptions. Indeed, several investigations have shown that time-related aspects (e.g., up-to-date data) are particularly important in data quality management (Klein and Callahan 2007, Sidi et al. 2012).

Despite their relevance for theory and practice, however, there is still a lack of well-founded and applicable data quality metrics to assess the currency of data in IS. Therefore, we state the following research question:

How should a metric be defined to assess the currency of data in IS?

To contribute to this question, we propose a probability-based currency metric (PBCM). By means of this metric, information about the currency of the assessed data can be considered in decision-making and add value in terms of better decisions. Indeed, the PBCM can also be seen as a possible basis for integrating data quality aspects in the theoretical framework of the value of information and particularly its probability-based normative concept (Carter 1985, Hilton 1981, Lawrence 1999, Marschak et al. 1972, Repo 1989).

The remainder of the paper is organized as follows. Section 2 illustrates the problem context and provides an overview of prior works. In Section 3, we develop a metric that is based on probability theory. The evaluation in Section 4 consists of two steps. First, we instantiate the metric for a real-use situation at a mobile services provider to demonstrate both its applicability and practical benefit. Second, we use publicly available real world data to demonstrate feasibility by defining an instantiation of the metric and to evaluate its strength. Finally, we summarize, reflect on the results, and provide an outlook on future research.

2. BACKGROUND

First, we provide some basic definitions and present the problem context. We then discuss existing contributions with respect to assessing the data quality dimension currency and identify the research gap.

2.1 Basic Definitions and Problem Context

Parssian et al. (2004, p. 967) use the terms information quality and data quality to "characterize mismatches between the view of the world provided by an IS and the true state of the world" (for a similar definition cf. Orr

1998). We take this definition as a basis. Data quality is a multi-dimensional construct (Lee et al. 2002, Redman 1996) comprising several dimensions such as accuracy, completeness, currency, and consistency (for an overview cf. Wang et al. 1995). Each dimension provides a particular view on the quality of attribute values in IS. We focus on currency and investigate how to assess this dimension by means of a metric.

Due to its relation to accuracy, we briefly discuss this data quality dimension in a first step. Afterwards, we define currency and delimit it from accuracy. Many authors (e.g., Batini and Scannapieco 2006, Redman 1996) define accuracy as the closeness of an attribute value stored in an IS to its real world counterpart. Usually, comparison or distance functions are used to determine the closeness of the attribute value with respect to its real world counterpart (Batini and Scannapieco 2006). The assessment of accuracy involves a *real world test* that constitutes a direct evaluation, for example by means of a survey or interview (cf. e.g., Wang and Strong 1996). Thus, both the stored attribute value and its real world counterpart are known when assessing accuracy.

In contrast to the widely accepted definition of accuracy, the definitions of time-related data quality dimensions are much less uniform in the literature (cf. Batini and Scannapieco 2006, Batini et al. 2009, Chayka et al. 2012). To express and specify time-related aspects, a number of different terms are used such as currency, timeliness, staleness, up-to-date, freshness, temporal validity, etc. Some contributions use different terms to define very similar or equal concepts while others use the same term describing different concepts. Ballou et al. (1998), for instance, refer to currency as the age of an attribute value at the instant of assessment. They use the term timeliness to describe whether "the recorded value is not out of date" (Ballou et al. 1998, p. 153). In contrast, Batini and Scannapieco (2006, p. 29) highlight that "currency concerns how promptly data are updated". Other authors such as Redman (1996, p. 258) state that currency "refers to a degree to which a datum in question is up-to-date. A datum value is up-to-date if it is correct in spite of possible discrepancies caused by time-related changes to the correct value". A similar definition is proposed by Nelson et al. (2005). Cho and Garcia-Molina (2003, p. 3) address an analog concept but use the term up-to-date to express that previously stored values "equal those of their real-world counterparts". Xiong et al. (2008, p. 952) also refer to a similar concept as that discussed by Redman (1996) and Nelson et al. (2005) but use the terms fresh and freshness, stating that "a real-time data object is fresh (or temporally valid) if its value truly reflects the current status of the corresponding entity in the system environment". This brief discussion illustrates that there is no widely accepted definition of such time-related data quality dimensions. As we primarily build upon the definitions of Redman (1996) and Nelson et al. (2005), we also use the term currency and clearly define the concept behind it for our context.

At its heart, currency expresses whether an attribute value that was stored in an IS in the past is still the same as the value of that attribute in the real world at the instant of assessment (i.e., in the present). This means that the attribute value, which was accurate when it was initially captured (Scenario A), updated (Scenario B), or acknowledged (Scenario C), is still the same as the current value of that attribute in the real world at the instant when its data quality is assessed. Currency explicitly focuses on the temporal decline of a stored attribute value. To illustrate this focus, we clarify the Scenarios A to C (cf. Figure 1).

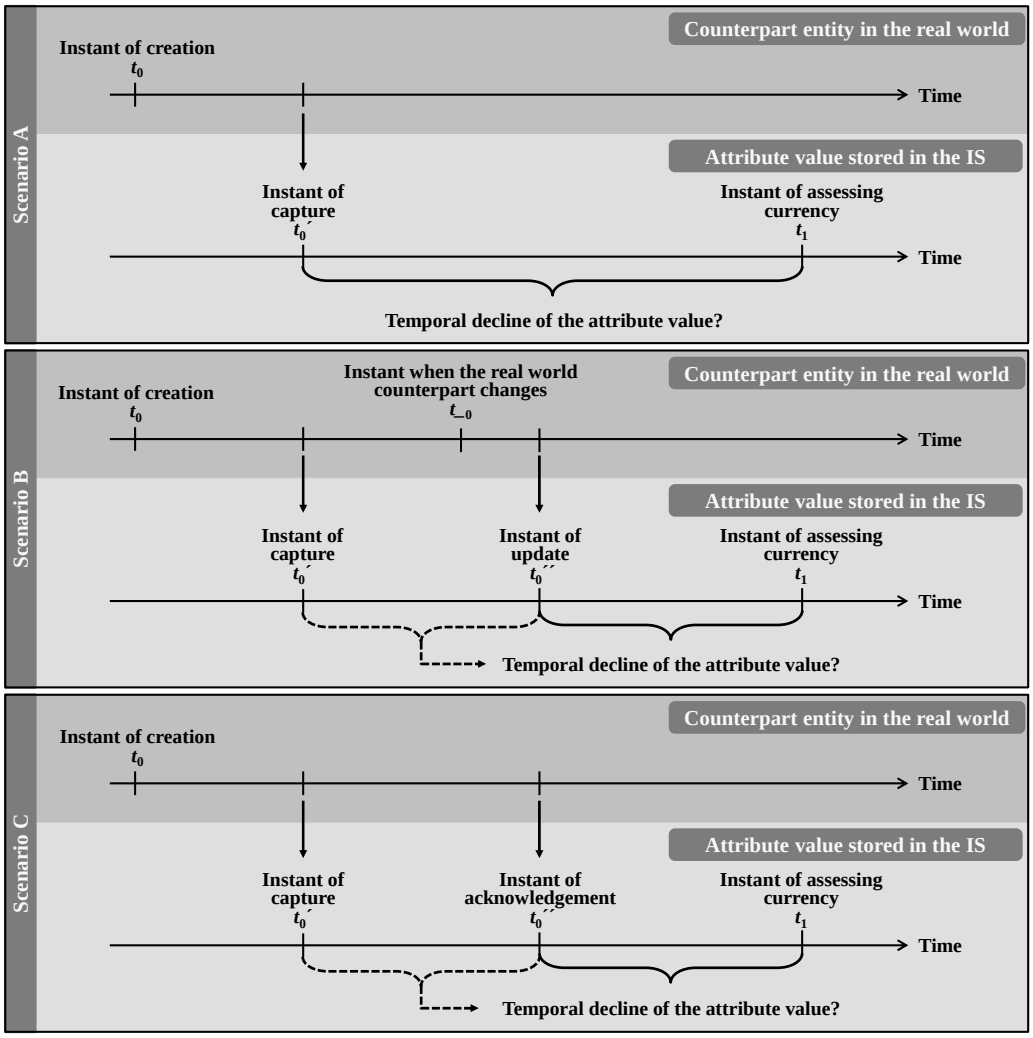

Fig. 1. Important Scenarios when Assessing the Currency of an Attribute Value

Scenario A shows the basic case: An attribute value was initially captured at the instant t_0' (i.e. the accurate value was stored in the IS). The instant of creation of its real world counterpart is represented by t_0. Here, it must be assessed whether the stored attribute value is still the same as the value of that attribute in the real world at the instant of assessment t_1. Thus, the question arises whether the real world counterpart has changed (which is unknown) since the attribute value was captured at t_0'. In Scenario B, it is known that the real world counterpart changed (e.g., at the instant $t_{\_0}$). The stored attribute value was therefore updated accordingly at the instant t_0''. In Scenario C, the stored attribute value was acknowledged at t_0'', as no changes had been made to

the real world counterpart. Both additional known instants (update and acknowledgement) can be useful (see below) when assessing the currency of the attribute value at the instant t_1.

When assessing the accuracy at the instant t_1, a real world test is needed. The result represents a statement under certainty. In contrast to assessing accuracy, assessing currency does not involve a real world test. Instead, a metric for currency delivers an indication, not a verified statement, as to whether an attribute value has changed in the real world since the instant it was captured, updated, or acknowledged.

An assessment of currency seems to be helpful in the following settings:

(a) *Unknown shelf life of the considered attribute value:* Assessing currency is helpful if the shelf life of the considered attribute value is unknown. Otherwise, the attribute value's currency can trivially be determined under certainty. The shelf life is defined as the length of time the stored attribute is still the same as the value of that attribute in the real world. In Scenario B, for example, the attribute value was created at the instant t_0 and changed at the instant t_0. Hence, the attribute value's shelf life was t_0-t_0. Possible application settings include customer master data such as name, address, phone number, marital status, number of children, profession, educational background, employer, and income, as the shelf life of the respective stored attribute values is usually unknown. However, even in the case of real world objects with a rather fixed shelf life such as credit cards that have a validity period of two years, for example, the shelf life of single attribute values may be unknown as the credit cards can become invalid earlier due to events such as withdrawal, theft, or loss of creditworthiness. Therefore, assessing currency can even be helpful in such cases. In addition to customer master data, the shelf life of product data, transaction data, and project data may also be unknown, which leads to further promising fields of application. The processes in production planning, for example, are typically based on data from a variety of internal and external sources (e.g., from suppliers or manufacturing partners) with the results strongly depending on the quality of the data used. In this context, the shelf life of the attribute values is unknown as well and assessing currency can provide helpful indications of whether the stored data are still the same as in the real world.

(b) *Real world test not possible or time-consuming or cost-intensive:* In the event the shelf life of an attribute value is unknown (cf. (a)), one can propose a real world test that directly compares the stored attribute value and the value of that attribute in the real world. However, such a real world test is often not practicable or too time-consuming and cost-intensive, for example, when customers have to be surveyed. For instance, analyses of data from a firm with more than 20 million customers show that every year about 2 million cus-

tomers change their place of residence, 230,000 die, and 60,000 get divorced (Schönfeld 2007). In this case, it would be very cost-intensive and impractical to regularly survey all customers in order to assess accuracy and update or acknowledge the stored attribute values. However, ignoring such data quality defects results in outdated data causing an annual loss of more than EUR 2 million for the firm due solely to inadequate customer contacts (Franz and von Mutius 2008). In other settings, for instance in distributed systems (e.g., in supply chains involving different firms), a real world test is hardly possible either. Therefore, it seems promising to assess currency and draw on indications whether a stored attribute value is still the same as the value of that attribute in the real world.

2.2 Related Work and Research Gap

A number of well-known and important contributions have been made with respect to the assessment of data quality (e.g., Batini et al. 2009, Batini and Scannapieco 2006, English 1999, Lee et al. 2002, Pipino et al. 2002, Redman 1996). In this subsection, we provide an overview of works on concrete metrics for assessing currency. One of the first and most renowned contributions was provided by Ballou et al. (1998). They define their metric[1] as a function depending on the age of the attribute value at the instant of assessing currency (t_1-t_0), the (fixed and given) shelf life of the attribute value, and a sensitivity parameter to adapt the metric to the context of application. The values of the metric are mapped on the interval [0; 1], with a value of one representing perfectly good and a value of zero perfectly bad currency. Hinrichs (2002) defines a metric for currency providing values normalized to the same interval. The values of this metric depend on how often the attribute values are (approximately) updated on average in the real world and the length of time between the instant of assessing currency and the instant of capturing the attribute value (t_1-t_0') or updating or acknowledging it (t_1-t_0''), respectively. Another metric is presented by Even and Shankaranarayanan (2007) in terms of a function depending on the length of time between the instant of assessing currency and the instant of capturing the attribute value (t_1-t_0') or updating or acknowledging it (t_1-t_0''). The main idea of their approach is to define the metric in such a way that its values denote the utility (represented by the interval [0; 1]) resulting from the currency of an attribute value. Two examples of utility functions (either the utility declines exponentially or completely when a certain threshold is reached) are discussed. Li et al. (2012) present a metric for currency in pervasive environments. This metric is defined based on the attribute value's storage age with respect to its last update (t_1-

[1] Ballou et al. (1998) actually propose a metric for the data quality dimension timeliness. Their understanding of timeliness, however, is comparable to the definition of currency consulted in this paper (cf. above).

t_0'') and its shelf life. To represent the update dynamics of pervasive data sources, the authors use the volatility in terms of the probability that another update happened since the last update (i.e., between t_0'' and t_1) as an exponent of an exponential function ("scaling factor"), which is added in a multiplicative way. Heinrich and Hristova (2014) present a metric based on expert estimations. Their metric is modelled as a fuzzy inference system consisting of a set of parallel IF-THEN rules. In this context, the authors also provide methods to estimate the input parameters for their metrics (i.e., age of the attribute value and its decline rate) by experts.

Another idea is to define currency based on probability theory. Heinrich et al. (2009) propose a procedure to develop probability-based metrics but explicitly do not seek to provide a concrete, mathematically noted one. Their procedure consists of six generic development steps – from selecting the attribute to be considered in the assessment via the identification of the impact factors that influence the shelf life of the respective attribute values through to defining and applying the metric. Heinrich et al. (2007, 2012) as well as Heinrich and Klier (2009, 2011) provide a formally defined metric. Assuming that the shelf life of attribute values is exponentially distributed, the values of the metric by Heinrich et al. (2007) and Heinrich and Klier (2011) can be interpreted as the probability that the attribute values are still current. Wechsler and Even (2012) propose a metric[2] based on a Markov-Chain model. Assuming memoryless transitions and an exponential probability distribution, this metric is similar to that of Heinrich et al. (2007) and Heinrich and Klier (2011). Obviously, however, the assumption of an exponential distribution does not hold for all attributes, which heavily affects the applicability of these approaches. Heinrich and Klier (2009) build on the idea of these probability-based approaches for attribute values characterized by an exponentially distributed shelf life. While providing first insights into how to consider supplemental data and an illustrative application scenario (using the metric to determine the Customer Lifetime Value), Heinrich and Klier (2009) as well as Heinrich et al. (2012), however, do not focus on missing or unknown supplemental data and a wide range of different data attributes and their specific characteristics.

Important contributions with respect to metrics for assessing currency in a wider sense are presented by Cappiello et al. (2003) as well as Pernici and Scannapieco (2003). Cappiello et al. (2003) provide insights into time-related dimensions of data quality in multichannel IS. They define mathematical functions to represent the currency of data on the level of operational databases as the average fraction of data that have not been modified or deleted in the interim in another operational database. Representing or assessing the currency of single

[2] They actually propose a metric for accuracy degradation. However, as Wechsler and Even (2012, p. 1) "observe accuracy and currency as related issues" and "address accuracies that are caused by failures to update data even when changes in the real-world entity require us to do so", their approach is well within our scope.

attribute values and assessing currency with regard to possible changes to the data in the real world, however, is beyond the focus of their work. Pernici and Scannapieco (2003) propose a data model and a methodological framework to associate quality information with data in web IS. They define a mathematical function to represent the volatility of data in terms of the temporal dynamics of the expiration of data. Their function corresponds to the probability that the expiration time associated with the data will change in the time interval starting from the instant of publication and ending with their expiration time. They do not aim to define a metric to assess the currency of attribute values in our sense; rather, the authors assume that the expiration time of data is given and stored in the IS when publishing the data, and try to declare the quality of the published data in terms of possible future changes and updates from the time the volatility is quantified onwards.

In summary, important contributions have been made with respect to metrics for currency. Compared to other approaches, a probability-based metric has some advantages. For instance, the metric values in terms of probabilities have a concrete unit of measurement and are interval scaled. However, there is still a research gap regarding a probability-based metric to assess currency that

(1) can cope with *missing or unknown metadata* (e.g., unknown shelf life and unknown instant of creation t_0 of the attribute value in the real world) for any number of attribute values,

(2) is able to cope with a *wide range of different data attributes* and their specific *characteristics* (e.g., changing decline rates) and does not depend on limiting assumptions (e.g., exponential distribution),

(3) is *formally modeled* (e.g., what are the exact underlying assumptions?) and *mathematically defined* (e.g., how can the metric be instantiated and how can its values be calculated?),

(4) takes into account *additional data* (e.g., other attribute values that are characterized by a statistical association with the considered attribute value's shelf life) to improve the strength of the metric, and

(5) has been *rigorously evaluated* (e.g., using publicly available real world data).

In the following section, we aim to develop and evaluate a probability-based metric that fills this gap and allows for a *reproducible configuration* and a *high level of automation* when assessing currency.

3. DEVELOPMENT OF THE PROBABILITY-BASED CURRENCY METRIC

In this section, we outline the conceptual foundations of our approach. On this basis, we develop the basic model of the PBCM. To improve the strength of the metric and to be able to cope with further realistic cases we then provide important extensions. Finally, possible ways to design instantiations of the PBCM are discussed.

3.1 Conceptual Foundations

Currency expresses whether an attribute value ω is still the same as the value of that attribute in the real world at the instant of assessment. In many contexts, the metadata of a real world counterpart, which means its shelf life and sometimes even its instant of creation, are unknown. Therefore, a metric for currency usually delivers an indication or estimation rather than a verified statement. We argue that the principles and the knowledge base of probability theory are adequate and valuable, providing well-founded methods to describe and analyze situations under uncertainty. Developing our metric, we interpret currency as the probability that an attribute value ω is still the same as the value of that attribute in the real world at the instant of assessing currency t_1 and has not become outdated due to temporal decline. In case of a limited and unknown shelf life, this probability decreases over time. If the shelf life of an attribute value ω is unlimited, the attribute value does not become outdated. This case is trivial and does not require an assessment. In addition, in contrast to some existing approaches we generally do not assume a fixed and known maximum shelf life of the attribute values because many attribute values either do not have a fixed maximum shelf life or it is not known.

Defining the metric values as a probability has several advantages: (1) Representing them as a probability ensures a concrete *unit of measurement* of the metric values; this means that the values of the metric are unambiguously defined and interpretable (Bureau International des Poids et Mesures 2006). (2) It seems natural and reasonable because a metric for currency delivers an indication or estimation under uncertainty. (3) The values of the metric are *interval scaled*, which means that the metric values as well as changes and differences in these values are meaningful (Frank and Althoen 1994). An interval scaled metric is necessary to compare, for instance, the effects of two or more data quality measures with each other and to interpret the resulting difference(s). (4) The metric values in terms of probabilities can be integrated, for instance, into the *calculation of expected values* to evaluate decision alternatives and support decision-making. Thus, each measure's data quality improvement can be compared to its costs to find out which measure is economically worthwhile.

3.2 Development of the Basic Model

Our basic model is based on the following assumptions and definitions (see Appendix A for an overview of all symbols and the mathematical notation used):

A.1 An attribute value ω is characterized by its real world counterpart's instant of creation t_0, which is initially known. The shelf life $T \in R^+$ of the attribute value ω is limited and unknown. It is regarded as stochastic (continuous random variable). The instant of assessing currency is represented by t_1 (with $t_1 \geq t_0$).

The age $t \in R^+$ of the attribute value ω can be determined by means of the instant of assessing currency t_1 and the instant of creation t_0 of the real world counterpart: $t=t_1-t_0$. An attribute value ω is current if and only if it is still the same as the value of that attribute in the real world at the instant of assessing currency t_1. This is the case, if and only if its shelf life T is greater than or equal to its age t. Because the shelf life T is unknown and is therefore regarded as stochastic, the currency of the attribute value ω cannot be determined under certainty. Consequently, currency is defined as the probability that the shelf life T is greater than or equal to the age t.

To support this currency assessment, additional data are considered. These additional data are characterized by a statistical association with the unknown shelf life T and therefore allow conclusions to be drawn about the shelf life T and the currency of the attribute value ω. To illustrate the relevance of additional data, we focus on the attribute value "student". Figure 2 illustrates the existence of a statistical association between the duration of study (including dropouts) – i.e., the shelf life T of the stored attribute value "student" – and the type of university (university vs. university of applied sciences). The cumulative frequency distributions are based on data provided by the Federal Statistical Office of Germany (2005-2008) and the Higher Education Information System GmbH (Heublein et al. 2003, 2008). Referring to persons whose data were captured in an IS when they started their study of Mathematics and Natural Sciences ten semesters ago: If it is additionally known that these persons enrolled at a *university of applied sciences*, it is expected that approximately 78% of them have already finished their studies and that the stored attribute value "student" is still current for only approximately 22% of them. If the persons enrolled at a *university*, however, it is expected that approximately 49% of the respective attribute values are still current. Thus, additional data – such as the type of university – seem to be relevant when assessing currency.

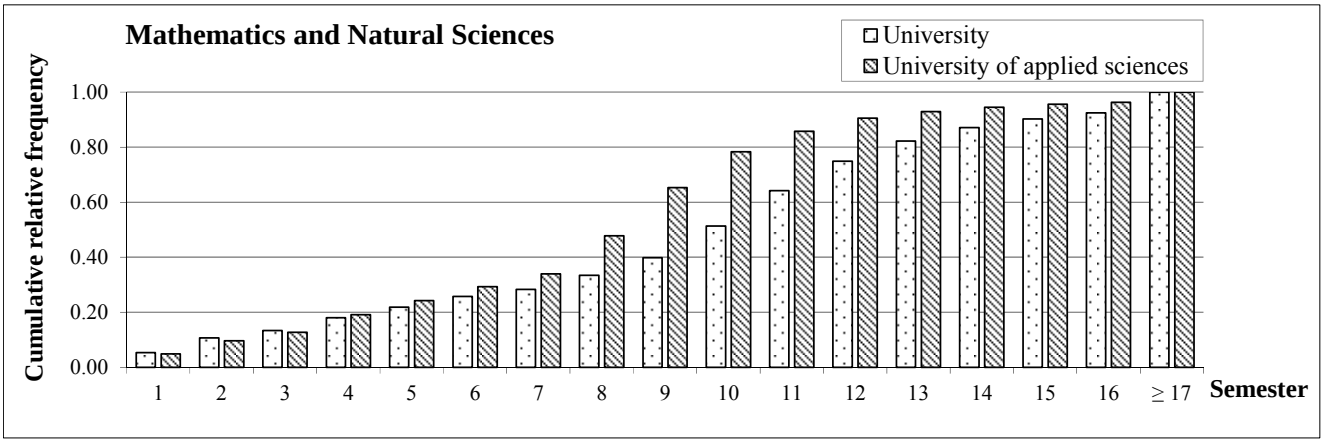

Fig. 2. Cumulative Frequency Distributions of the Duration of Study (incl. Study Dropout)

Assumption A.2 takes this aspect into account. Moreover, it also reflects the fact that in real databases parts of

the relevant additional data may not be stored (i.e., they are not known) for all attribute values ω.

A.2 The cumulative distribution function $F^{\omega}(t|w_1,...,w_n) := P^{\omega}(T \leq t | W_1 = w_1,...,W_n = w_n)$ of the shelf life T of attribute value ω is given. It depends on the additional data w_i (with $i=1,\ldots,n$) being part of the data record. The additional data w_i represent realizations of the random variables W_i. At the instant of assessing currency t_1 – without any loss of generality – only the additional data w_j (with $j=1,\ldots,l$ and $l \leq n$) are known.

Based on assumptions A.1 and A.2, we define the PBCM for the attribute value ω as the *conditional probability* that the shelf life T of the attribute value ω is greater than or equal to its age t. The known additional data w_j (with $j=1,\ldots,l$ and $l \leq n$) serve as conditions $W_1=w_1, \ldots, W_l=w_l$. The additional data w_k (with $k=l+1,\ldots,n$) are unknown for the attribute value ω. Hence, currency has to be assessed without knowing the realizations of the random variables W_k. Because these realizations are part of the distribution function $F^{\omega}(t|w_1,...,w_n)$ (see A.2), we use an expected value calculus in order to remove them from the corresponding density function $f^{\omega}(\theta|w_1,...,w_n)$. This can be done by integrating the density function over the universal sample spaces Ω_{W_k} of the random variables W_k (with $k=l+1,\ldots,n$) (cf. proof in Appendix B; note that $\int_{\Omega_{W_n}} \ldots \int_{\Omega_{W_{l+1}}} f^{\omega}(w_1,...,w_n)\,dw_{l+1}...dw_n > 0$ is necessary):

$$f^{\omega}(\theta \mid w_1,...,w_l) = \frac{\int_{\Omega_{W_n}} \ldots \int_{\Omega_{W_{l+1}}} f^{\omega}(\theta, w_1,...,w_n)\,dw_{l+1}...dw_n}{\int_{\Omega_{W_n}} \ldots \int_{\Omega_{W_{l+1}}} f^{\omega}(w_1,...,w_n)\,dw_{l+1}...dw_n} \tag{1}$$

We define the PBCM $Q^{\omega}_{Curr.}(t, w_1,...,w_l)$ as denoted in Term (2):

$$\begin{aligned} Q^{\omega}_{Curr.}(t, w_1,...,w_l) &:= P^{\omega}(T \geq t \mid W_1 = w_1,...,W_l = w_l) = \\ &= 1 - P^{\omega}(T < t \mid W_1 = w_1,...,W_l = w_l) = 1 - \int_0^t f^{\omega}(\theta \mid w_1,...,w_l)\,d\theta \end{aligned} \tag{2}$$

The PBCM is defined based on the complementary probability $P^{\omega}(T<t|W_1=w_1,...,W_l=w_l)$, which represents the probability that the attribute value is outdated at the instant t_1 ($T<t=t_1-t_0$) considering the known additional data w_j (with $j=1,\ldots,l$ and $l \leq n$) as conditions. As the complementary probability represents whether the attribute value ω has become outdated before the age t is reached, the definite integral of the determined density function $f^{\omega}(\theta|w_1,...,w_l)$ is calculated over the interval [0; t]. In contrast to existing approaches, our approach allows us to consider additional data in a well-founded way. Using the expected value calculus according to Term (1) ensures that even in the case of missing additional data the metric can be applied in a widely automated way and

does not require any further manual configuration. To illustrate this expected value calculus, we use the example introduced in Figure 2. In the event the additional data regarding the type of university is unknown when assessing the currency of the attribute value "student" of a specific person, all possible values of the additional data (realizations of the random variable type of university) and their probabilities have to be taken into account. Hence, we do not consider only one single realization when assessing currency, but rather integrate each possible realization of the random variable with its corresponding probability into our calculation (cf. Term (1)). In the example, the probabilities that the person enrolled at a university and a university of applied sciences, respectively, are considered. However, it has to be noted that the strength of the metric is affected if additional data are unknown. A mathematical discussion of this fact is provided in Appendix C.

As the shelf life T is limited, the decline rate $z^{\omega}(t|w_1,...,w_l)$ is a key characteristic of $P^{\omega}(T\leq t|W_1=w_1,...,W_l=w_l)$ and therewith the cumulative distribution function $F^{\omega}(t|w_1,...,w_n)$. By multiplying this decline rate with the length h of the time interval $[t; t+h]$ (with $h\in R^+$), it is possible to approximate the probability that the attribute value ω which is still current at an age t becomes outdated in the following time interval of length h ($P^{\omega}(T\leq t+h|W_1=w_1,...,W_l=w_l,T>t)$). Based on Term (1), $z^{\omega}(t|w_1,...,w_l)$ can be represented as follows:

$$
\begin{aligned}
z^{\omega}(t \mid w_1,..., w_l) &:= \lim_{h\to 0}\left(\frac{P^{\omega}(T \leq t+h \mid W_1 = w_1,..., W_l = w_l, T > t)}{h}\right) = \\
&= \lim_{h\to 0}\left(\frac{P^{\omega}(T \leq t+h \mid W_1 = w_1,..., W_l = w_l) - P^{\omega}(T \leq t \mid W_1 = w_1,..., W_l = w_l)}{h \cdot P^{\omega}(T > t \mid W_1 = w_1,..., W_l = w_l)}\right) \\
&= \frac{f^{\omega}(t \mid w_1,..., w_l)}{1 - \int_0^t f^{\omega}(\theta \mid w_1,..., w_l)\,d\theta}.
\end{aligned}
\tag{3}
$$

The continuous decline rate $z^{\omega}(t|w_1,...,w_l)$ is defined as the limit of the quotient of the conditional probability $P^{\omega}(T\leq t+h|W_1=w_1,...,W_l=w_l,T>t)$ and the length h of the time period $[t; t+h]$ as h approaches zero. Multiplying the decline rate $z^{\omega}(t|w_1,...,w_l)$ with h can only serve as an approximation of $P^{\omega}(T\leq t+h|W_1=w_1,...,W_l=w_l,T>t)$. This is because the decline rate may be piecewise constant, increasing and/or decreasing depending on $F^{\omega}(t|w_1,...,w_n)$. Moreover, by using the decline rate $z^{\omega}(t|w_1,...,w_l)$ it is possible to (intuitively) represent the specific characteristics of data attributes in terms of constant, increasing, decreasing, or changing decline rates.

3.3 Extensions of the Basic Model

In the following subsection, we provide extensions of the basic model to improve the strength of the PBCM and have the ability to address realistic cases when (additional) metadata are available or unknown.

3.3.1 Considering Additional Metadata Referring to Data Updates or Acknowledgements

In Scenarios B and C, the PBCM has to deliver an indication whether a stored attribute value ω, which was updated or acknowledged in the meantime, is still current. In the following, we extend the basic model to take into account such additional metadata distinguishing the cases of (a) updates and (b) acknowledgements.

(a) According to Scenario B, attribute value ω constitutes an update of a formerly stored attribute value ω' at the instant t_0''. This update is due to a change of the real world counterpart at the instant t_0 (cf. Figure 1). When assessing currency, the formerly stored attribute value ω' may serve as additional metadata. Such data are usually available in case of temporal databases and bitemporal timestamps. Indeed, it is possible to observe former changes of the attribute value over time, as well as the lengths of time the attribute value was stored. To take these additional metadata into account the PBCM has to be extended. Here, the additional metadata serve as further conditions apart from the conventional additional data. Term (4) shows the extended PBCM considering the formerly stored attribute value ω' (the respective random variable is denoted as $W_{\omega'}$) when assessing the currency of the attribute value ω with age $t=t_1-t\_0$:

$$Q^{\omega}_{Curr.}(t,\omega',w_1,...,w_l) := P^{\omega}(T \geq t \mid W_{\omega'} = \omega', W_1 = w_1,..., W_l = w_l) = \\ = 1 - \int_0^t f^{\omega}(\theta \mid \omega', w_1,..., w_l)\,d\theta \tag{4}$$

Considering the formerly stored attribute value ω', the strength of the metric can be improved (this can be proven analogously to the case of conventional additional data; cf. Appendix C). The relevance of considering such additional metadata can be illustrated by using the attribute value "student". The average dropout rates depend heavily on a person's former school education and are much higher for persons who graduated from an evening school or a vocational college compared to those who attended a regular secondary school. Considering the additional metadata referring to the previous school education of a person significantly affects the probability of that person still being a student and, consequently, the currency.

(b) According to Scenario C, the attribute value ω was acknowledged at the instant t_0'' (cf. Figure 1). One could believe that knowing about the instant of acknowledgement is worth nothing when assessing currency. However, it can easily be shown that considering the instant t_0'' as additional metadata can significantly improve the strength of the PBCM. We illustrate this fact using the example of the attribute value "student". Referring to persons whose data were captured when they started their studies of Mathematics and Natural Sciences at a university of applied sciences ten semesters ago, it is expected that the stored at-

tribute value "student" is still current for only approximately 22% of them (cf. Figure 2). If it is additionally known that the attribute value "student" was acknowledged one semester ago (at the instant t_0''), it is expected that the attribute value "student" is still current for approximately 63% (=(1-0.78)/(1-0.65)) of them (cf. Figure 2). Additional metadata regarding the instant t_0'' are obviously relevant. Hence, in Term (5) we extend the basic model by considering the age t' in terms of the length of time between the instant when the attribute value ω was acknowledged t_0'' and its instant of creation t_0 (i.e., $t'=t_0''-t_0$).

$$Q^{\omega}_{Curr.}(t,t',w_1,...,w_l) := \frac{P^{\omega}(T \geq t \mid W_1 = w_1,...,W_l = w_l)}{P^{\omega}(T \geq t' \mid W_1 = w_1,...,W_l = w_l)} = \frac{1-\int_0^t f^{\omega}(\theta \mid w_1,...,w_l)d\theta}{1-\int_0^{t'} f^{\omega}(\theta \mid w_1,...,w_l)d\theta} = \tag{5}$$

$$= \frac{Q^{\omega}_{Curr.}(t,w_1,...,w_l)}{Q^{\omega}_{Curr.}(t',w_1,...,w_l)}.$$

The contribution of this extension does not only stem from the fact that an enhanced solution is provided for relevant cases. In fact, the metric as defined in Term (5) does not require another cumulative distribution function but can be traced back to applying the basic model as defined in Term (2) for t and t'.

3.3.2 Addressing an Unknown Instant of Creation t_0

According to assumption A.1, the instant of creation t_0 is known for all attribute values. This may be unrealistic as an attribute value's instant of creation t_0 may be stored only sporadically when capturing the attribute value. Therefore, we will address the case of an unknown instant of creation t_0.

Many databases store the instants when attribute values were captured, updated, or acknowledged as metadata (e.g., "last modified" attribute). In the following, the latest of these instants will be referred to as the instant of data entry t_0^* (with $t_0 \leq t_0^* \leq t_1$). In case the instant of creation t_0 is unknown, the known instant of data entry t_0^* is used to assess currency. Here, the age $t^* \in R^+$ of the attribute value ω with respect to the instant of data entry t_0^* (i.e., the storage time) can be determined to $t^*=t_1-t_0^*$. Thus, based on t^* and the cumulative distribution function $F^{*\omega}(t^*|w_1,...,w_n) := P^{\omega}(T^* \leq t^* | W_1=w_1,...,W_n=w_n)$ of the shelf life T^*, we define the metric for currency $Q^{*\omega}_{Curr.}(t^*,w_1,...,w_l)$ as denoted in Term (6). Again its values represent the probability that the attribute value ω is still current at the instant of assessing currency t_1. The density function $f^{*\omega}(\theta|w_1,...,w_l)$ can be determined as in Term (1) based on $F^{*\omega}(t^*|w_1,...,w_n)$ and the corresponding density function $f^{*\omega}(\theta|w_1,...,w_n)$.

$$Q^{*\omega}_{Curr.}(t^*,w_1,...,w_l) := P^{\omega}(T^* \geq t^* \mid W_1 = w_1,...,W_l = w_l) = 1 - \int_0^{t^*} f^{*\omega}(\theta \mid w_1,...,w_l)d\theta \tag{6}$$

Thereby, the case of the distribution $F^{\omega}(t|w_1,...,w_n)$ being a memoryless distribution is particularly interesting. Exponential and geometric distributions constitute memoryless probability distributions and play an important role in data quality assessment. For instance, the exponential distribution is frequently discussed in the context of address data (cf. e.g., Heinrich et al. 2007). In the case of a memoryless probability distribution, the attribute value is characterized by a constant, relative decline rate $z^{\omega}(t|w_1,...,w_n)$. Hence, the distribution functions $F^{\omega}(t|w_1,...,w_n)$ and $F^{*\omega}(t^*|w_1,...,w_n)$ and therewith $P^{\omega}(T \leq t | W_1=w_1,...,W_l=w_l)$ and $P^{\omega}(T^* \leq t^* | W_1=w_1,...,W_l=w_l)$ are equal. Considering either the shelf life T and the age t (within the basic model) or the shelf life T^* and the storage time t^* (within the extended model) makes no difference and yields the same result:

$$\begin{aligned} &P^{\omega}(T \geq t \mid W_1 = w_1,...,W_l = w_l, T > t - t^*) = P^{\omega}(T \geq t^* \mid W_1 = w_1,...,W_l = w_l) = \\ &= 1 - P^{\omega}(T \leq t^* \mid W_1 = w_1,...,W_l = w_l) = 1 - P^{\omega}(T^* \leq t^* \mid W_1 = w_1,...,W_l = w_l) = Q^{*\omega}_{Curr.}(t^* \mid w_1,...,w_l) \end{aligned} \tag{7}$$

Hence, if we identify an (approximately) constant, relative decline rate $z^{\omega}(t|w_1,...,w_n)$ when determining $F^{\omega}(t|w_1,...,w_n)$, a memoryless distribution can be assumed. In that case, it is not necessary to know the instant of creation t_0 of an attribute value ω; the instant of data entry t_0^* is sufficient to yield the same result.

3.4 Possible Ways to Develop Instantiations of the Metric

To instantiate the PBCM, it is necessary to determine the cumulative distribution function of the shelf life of the attribute values $F^{\omega}(t|w_1,...,w_n)$, and, in this context, especially the corresponding attribute-specific decline rate $z^{\omega}(t|w_1,...,w_n)$. To do so, we scratch the following possibilities:

1. Analysis of publicly available data (e.g., from public or scientific institutions)
2. Analysis of company-owned (historical) data (e.g., from the data warehouse)
3. Conducting a study (e.g., surveying a sample of customers to determine the decline rate)
4. Surveying experts (e.g., determining the decline rate based on experts' estimations)

The first possibility refers to the use of publicly available data. Here, data about factors which influence the decline rate of the data attribute considered have to be acquired (e.g., from federal statistical offices, public or scientific institutions). The attribute value "student", for instance, can become outdated due to two main factors. A study is either completed or aborted. By means of publicly available data regarding both factors, the distribution function $F^{\omega}(t|w_1,...,w_n)$ can be determined (cf. next section). Attributes such as *last name*, *marital status*, and *address* may serve as further examples. The decline rates for *last name* and *marital status* can be determined using publicly available data regarding marriages and divorces. The same holds for *address* and publicly available data regarding the frequency of relocation.

If no such third party data are available, company-owned (historical) data may be analyzed. This kind of analysis seems favorable, if assessing currency concerns company-specific data attributes. To determine the decline rate of the attribute *current tariff*, for instance, historical customer data could be extracted from the company's operational databases or data warehouse. Based on this data extraction, the average duration of different contracts and tariffs can be calculated. Besides, the decline rate of the attribute *address* may also be determined based on company-owned data. If a company's customers are characterized by specific characteristics, the analysis of company-owned data may have advantages over the analysis of publicly available data that often do not take into account such specific characteristics (e.g., seniors typically have a lower frequency of relocation).

Conducting a study is a further possibility to determine attribute-specific decline rates. Focusing, for instance, on assessing currency of a customer-specific attribute (e.g., the attribute *employer* of a customer), a random sample of the customer base can be drawn. These customers could be surveyed to get data on the shelf life of the attribute values considered. Such data can be used to determine both the decline rate and the distribution function. A short example considering the attribute *employer*, which is particularly important for financial services providers offering financial planning products, may illustrate this. If data are needed to determine the frequency of job changes, it is possible to draw a sample of customers and survey them. After determining the validity period of an employment, the decline rate and the distribution function can be calculated.

Finally, decline rates based on experts' estimations can be used. This may be reasonable, if neither external nor internal data are available and conducting a study is too costly. These cases occur rarely because assessing currency mostly concerns attribute values stored in existing databases (thus historical data should be available). However, experts' estimations are still relevant. Here, the Delphi method may be used which is a systematic, interactive method that relies on a panel of experts. For example, instead of using historical data, a company's key account managers may be surveyed to determine the decline rate of the attribute *current tariff*.

After determining the distribution function and the corresponding decline rate which has to be done only once for each attribute considered, it is possible to develop instantiations of the PBCM (cf. next section). Then, the values of the metric can be calculated for all considered values of an attribute. Here, the input parameters of the metric in terms of available instants of creation, update, or acknowledgement and additional data w_i are used for each attribute value ω. These input parameters can easily be extracted from the database and exported by means of SQL DML statements, provided they are stored. These data can be processed, for example, by a Java program that implements the PBCM and makes it possible to calculate its values for all attribute values.

4. EVALUATION OF THE PROBABILITY-BASED METRIC FOR CURRENCY

First, we demonstrate the applicability and the practical benefit of the PBCM by means of a case study. Second, the feasibility and the strength of the metric are analyzed based on publicly available data.

4.1 Evaluation of the Metric by Means of a Case Study

The goal of this evaluation step is to analyze the feasibility and the applicability as well as the practical benefit of the PBCM by means of a case study. The following evaluation questions are examined:

E.1 How can the PBCM be instantiated and applied in a real-use situation?

E.2 How can the values of the PBCM in terms of probabilities be integrated into a decision calculus?

E.3 What is the practical benefit resulting from the application of the PBCM?

4.1.1 Case Selection and Starting Point in the Case

The PBCM was applied in the campaign management at a major German mobile services provider. For reasons of confidentiality, figures and data had to be changed and made anonymous. Nevertheless, the procedure and the basic results remain unchanged. Both the sales & marketing and the data warehouse departments identified data quality issues in recent campaigns. These became especially apparent when – as a follow-up of a conducted campaign – randomly selected customers were asked about the reasons why they did not accept the offer. The survey showed that more than 33% of the selected customers no longer belonged to the focused target group of the campaign. This was due to outdated customer data that were used as selection criteria for the target group. Hence, a forthcoming student campaign was chosen to analyze whether and how the customer selection could be supported using the PBCM. The aim of this campaign is to offer customers with student status a new premium tariff called *ForStudents 500* by mail. For reasons of price differentiation, this tariff is only available for customers who are actually students. Hence, the attribute value "student" is used as a criterion for the customer selection of this campaign. In case the attribute value "student" is outdated, this customer cannot accept the offer (confirmation of enrollment required). As a result of outdated data customers who have already finished or abandoned their studies are included in the target group of the campaign. Selecting wrong customers, however, results in decreased customer satisfaction and low campaign success rates. To alleviate this problem, we initially analyzed the existing customer selection procedure for such campaigns:

1. All customers fulfilling the selection criterion (e.g., *professional status* = "student") were identified. In the case of the student campaign considered, approximately 170,000 customers were selected.
2. For each of these customers the previous year's sales volume was extracted because the higher the sales

volume of a customer accepting the offer the higher the additional returns resulting from the campaign. For the student campaign the additional return was estimated to be 6% of the customer's previous sales volume.

3. Based on their previous sales volume, the top X% customers were selected to constitute the target group. For the student campaign, the top 30% customers (i.e., 51,000 customers) were selected, which was a requirement of the marketing department (reasons of exclusivity and to strengthen customer loyalty).

In the past, the success rates of campaigns using this procedure averaged 9%. Thus, the number of customers accepting the offer was estimated at 4,590 (=9% · 51,000). The average sales volume of the top 30% customers was calculated to be EUR 1,470, resulting in an estimated additional return *r* per customer of EUR 88.20 on average (=EUR 1,470 · 6%). This was supposed to lead to a total additional return *R* of about EUR 404,838 (=4,590 · EUR 88.20) which was significantly higher compared to the estimated costs.

4.1.2 Adapted Selection Procedure and Instantiation of the Metric

To consider the values of the PBCM when selecting the target group, we defined an adapted procedure. This way, the decision was supported, for instance, regarding whether it made sense from an economic point of view to address a customer characterized by a low probability of still being a student who might therefore be unable to accept the offer. The customer database comprised approximately 170,000 customers with the attribute value "student". Hence, it was necessary to assess currency in a widely automated way and avoid a manual calculation of each single value of the metric. The adapted customer selection procedure was defined as follows:

1. All customers fulfilling the selection criterion (e.g., *professional status* = "student") were selected.
2. For each of these customers, the previous year's sales volume was extracted.
3. For each of these customers, the individual value of the metric was calculated.
4. The sales volumes and the values of the metric in terms of probabilities were integrated into an expected value calculus; for each customer the expected value of the additional return $E(r) = Q^{\omega}_{Curr.}(t, w_1, ..., w_n) \cdot r$ was calculated in an automated way and used as a criterion to identify the top 30% of customers.

The database of the mobile services provider comprised, among others, the attributes *instant of enrollment* (t_0), *type of university* (W_1), and *field of study* (W_2). With respect to the PBCM, the values of the last two attributes could serve as additional data. The values w_1 of the attribute *type of university* (W_1) included "University" and "University of applied sciences". The values w_2 of the attribute *field of study* (W_2) came from the list "Economics and Social Sciences", "Engineering Sciences", "Mathematics and Natural Sciences", "Law", "Agriculture, Forestry and Food Sciences", "Education/Teaching Sciences", "Linguistic/Philology and Cultural Sciences",

"Art", and "Health Sciences". These values were retrieved during data acquisition where possible. However, they were not stored for all customers (represented by NULL in the database).

When developing the PBCM for the attribute value "student", factors influencing its shelf life had to be determined. This attribute value can lose its currency due to two factors: a study is either completed successfully or aborted. Hence, the PBCM had to take into account both factors by means of the corresponding conditional probability distributions depending on the age t of the attribute value and the values of the random variables W_1 and W_2 (additional data). The conditional distribution of the shelf life T, measured in number of semesters, could be determined easily based on publicly available data provided by the Federal Statistical Office of Germany (2005-2008) and the Higher Education Information System GmbH (Heublein et al. 2003, 2008).

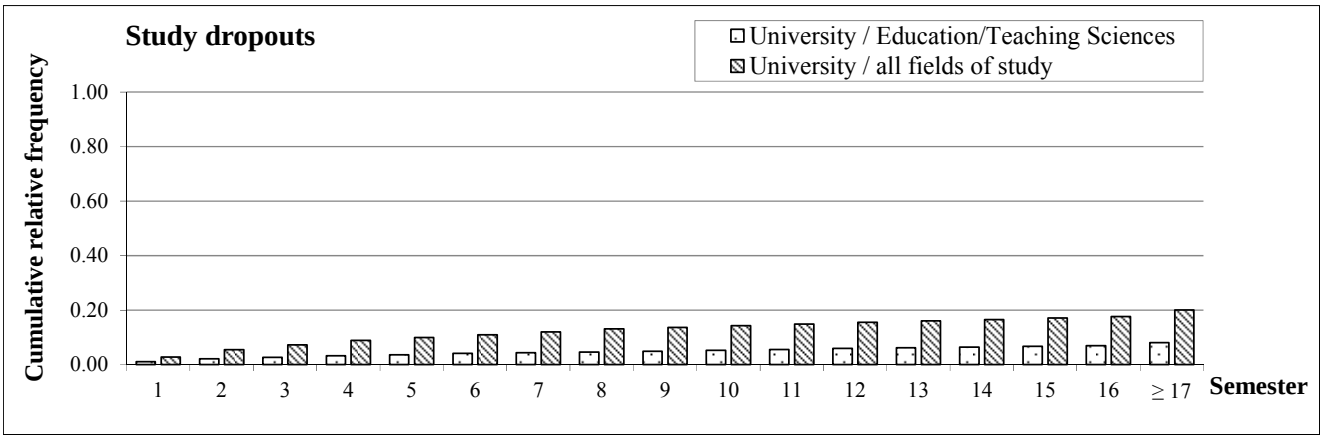

Fig. 3. Cumulative Relative Frequency Distributions of Study Dropouts

Figure 3 shows the cumulative relative frequency distributions of study dropouts at universities for Education/Teaching Sciences and over all fields of study. The frequencies depending on the *type of university* and the *field of study* could be calculated for each *type of university* by multiplying the fraction of study dropouts in the respective semesters with the corresponding overall dropout rate for the respective *field of study*. To be able to cope with unknown additional data it was necessary to calculate for each *type of university* the expected value regarding the *field of study*. Here, we used the weighted average of the cumulative relative frequencies over all possible values for the *field of study*. The fractions of the number of students in the particular *field of study* with respect to the overall number of students at the respective *type of university* served as weights. Based on this, it was possible to calculate the probability that a customer with the attribute value "student" had already dropped his or her studies after a duration of study of t semesters. The duration of study t was represented by the difference between the instant of assessing currency (t_1=start of the summer semester 2009) and the student's instant of enrollment t_0. This probability is referred to as the dropout probability $P(Dropout \leq t | W_1=w_1, \ldots, W_n=w_n)$.

In an analogous way, the distributions for a successful completion of study were determined. Figure 4 illustrates the cumulative relative frequency distributions for students enrolled at universities for Education/Teaching Sciences and over all fields of study. The conditional probability that a customer has already successfully completed his or her studies after *t* semesters is represented by $P(Graduate \leq t | W_1=w_1, ..., W_n=w_n)$.

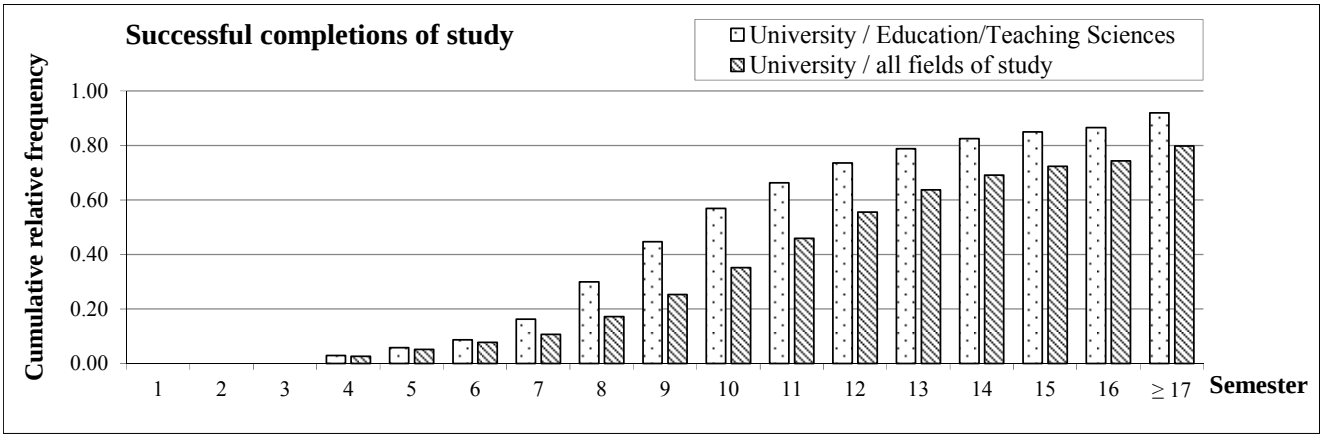

Fig. 4. Cumulative Relative Frequency Distributions of Successful Completions of Study

Based on the probabilities for dropout $P(Dropout \leq t | W_1=w_1, ..., W_n=w_n)$ and the successful completion $P(Graduate \leq t | W_1=w_1, ..., W_n=w_n)$ of studies the PBCM was instantiated as follows:

$$Q^{\omega}_{Curr.}(t, w_1, ..., w_n) := 1 - (P(Dropout \leq t \mid W_1 = w_1, ..., W_n = w_n) + P(Graduate \leq t \mid W_1 = w_1, ..., W_n = w_n)) \quad (8)$$

Table 1 shows four selected customers. For customer D the value of the metric is 0.38. For customer C the probability was calculated to 0.51, not knowing the *field of study*. Considering all possible values of the attribute *field of study* resulted in a maximum absolute error of 0.41 of the metric value for customer C, which illustrates the relevance of additional data; the expected absolute error was calculated to 0.12.

Customer	**Professional status**	**Instant of enrollment t_0**	**Type of university W_1**	**Field of study W_2**	$Q^{\omega}_{Curr.}(t, w_1, w_2)$
A	"student"	Summer semester 2004	"University"	"Mathematics and Natural Sciences"	0.49
B	"student"	Summer semester 2004	"University of applied sciences"	"Mathematics and Natural Sciences"	0.22
C	"student"	Summer semester 2004	"University"	NULL	0.51
D	"student"	Summer semester 2004	"University"	"Education/Teaching Sciences"	0.38

Table 1. Example of Four Selected Customers and their Metric Values

4.1.3 Application of the Metric and Results of the Ex Post Analysis

For each of the 170,000 customers, the metric value was calculated in an automated way. Then, the additional

return r in case the offer is accepted was determined for each customer. On this basis, the expected value $E(r)$ of the additional campaign return was calculated for each customer and used as a selection criterion to identify the top 30% of customers. Based on the new selection criterion, customers other than those in the existing campaign management procedure were selected. For instance, many customers with a relatively high sales volume but a very small probability of still being a student were now omitted. Overall, only 20,130 customers were selected according to both selection criteria (sales volume and expected additional return) while 30,870 customers were only selected based on one selection criterion. Therefore, the marketing department decided, as a precaution, to address all 81,870 customers selected according to at least one selection procedure.

Table 2 shows the results of the ex post analysis: The success rate observed for the 51,000 customers selected according to the existing procedure was approximately 7.7%. This value was lower compared to the expected success rate of 9%. Based on the average sales volume of EUR 1,450 of the customers who *actually* accepted the offer, the total additional ex post return R was EUR 342,780 (=3,940 · EUR 1,450 · 6%). Considering the customers selected according to the adapted procedure, we can state: on the one hand, the average sales volume of EUR 1,300 per customer who *actually* accepted the offer was lower compared to the existing procedure. On the other hand, the success rate of approximately 17.3% even exceeded the expected success rate of 9% and resulted in a much higher total additional ex post return R of EUR 687,180 (=8,810 · EUR 1,300 · 6%).

	Existing Selection Procedure	**Adapted Selection Procedure Based on the PBCM**
Number of customers with attribute value "student"	170,000	170,000
Number of customers selected for the campaign	51,000	51,000
Number of customers accepting the offer	3,940	8,810
Avg. sales volume of customers accepting the offer	EUR 1,450	EUR 1,300
Success rate	7.7%	17.3%
Additional return of the campaign R	EUR 342,780	EUR 687,180

Table 2. Results of the Ex Post Analysis

4.1.4 Further Applications of the Metric in Real-Use Situations

The PBCM has also been successfully applied in further real-use situations. The following brief discussion illustrates that the metric cannot only be used in the context of customer databases. Rather, it can be applied in different industries and business divisions, such as production planning and control, supply chain management, and controlling. Most notably, our metric has also been applied in the scenarios described below.

The metric was employed by a German automotive company to assess the currency of data stored in the global

supply chain management IS. This IS executes, among others, several hundred thousand spare part orders from dealers, dealer and delivery centers, country headquarters, and the central headquarter. A major problem is that the stored status for orders may be outdated. Already processed or canceled orders that are still stored as open may serve as examples. Differences between the stored and the real status of an order mostly occur due to the large number of users of the IS and the various manual data capturing and transformation processes. To keep the IS up-to-date and avoid problems when further processing the data, it is essential to assess the currency of the values of the attribute *order status*. The identification of outdated orders, however, is not easy because the actual processing time of orders usually varies from several days to several months depending on the specific order properties. Hence, it is necessary to consider additional data such as an order's priority, its delivery channel, and the country of the ordering organization. A calculation of the average processing time over all orders neglecting additional data does not make sense because outdated orders with special characteristics (e.g., such as specific countries) would be identified too late. Moreover, using the average processing time of orders, too many orders would be categorized as potentially outdated and overwhelm the analysis. In contrast, applying our metric considering additional data made it possible to determine the currency of the status of the spare part orders in terms of probabilities in a well-founded way. Based on the metric values, the company decided to examine only those orders that were outdated with a probability of more than 80%. In so doing, extensive analysis costs could be avoided, and the targeted identification of outdated orders significantly helped to reduce wrong invoicing or non-invoicing of orders. Moreover, the number of ordered spare parts that were delivered late or not at all could be reduced as well – which improves customer satisfaction and may prevent losing customers.

A case conducted in cooperation with a globally acting furniture manufacturing company focuses on a completely different context, namely the management of employees' project skills. These skills are documented in a database to support project staffing. However, the problem is that the stored skill profiles are outdated if employees have gained new or additional skills, or extended their knowledge or experience that has not been documented accordingly. This makes it difficult to successfully search for employees with specific skills needed to adequately staff projects. To ease this drawback, the PBCM was used to determine the probabilities that the skill profiles of employees had not become outdated. In this context, a simple calculation of the average period of time a stored skill profile remains up-to-date is not very helpful. This is due to the fact that the currency of a skill profile depends heavily on various additional data (e.g., job description, department, position, etc.) characterizing the individual employee. By means of the PBCM, however, the currency in terms of the probability

that a stored profile is still up-to-date could be calculated for each employee, helping to initiate, in an almost fully automated way, an update of the employee's skill profile, carried out by him-/herself or his/her supervisor (e.g., when the probability that the profile is outdated is above 75%). With selective calls for updates based on the metric values, inefficiencies were reduced and acceptance of the skill database was improved.

4.1.5 Results of the Evaluation Step

The case study demonstrates the feasibility and applicability of the PBCM and provides a real world situation in which the metric leads to considerable economic benefits. Table 3 sums up the results.

Evaluation Question	Result
E.1 How can the PBCM be instantiated and applied in a real-use situation?	We described how we used publicly available data to instantiate the PBCM for the case of the German mobile services provider (for further possible ways to develop instantiations of the metric cf. Section 3.4). We also illustrated how we successfully applied the PBCM in the field of campaign management. Indeed, by means of the PBCM the currency of attribute values could be taken into account in the customer selection procedure.
E.2 How can the values of the PBCM in terms of probabilities be integrated into a decision calculus?	Due to their unambiguous interpretation as probabilities, the integration of the values of the PBCM into a decision calculus could be done easily and in a well-founded manner (in contrast to most existing metrics). The metric values in terms of probabilities (currency indication) were used to calculate the expected value of the additional return of a customer (selection criterion).
E.3 What is the practical benefit resulting from the application of the PBCM?	We analyzed the economic effects of using our metric in campaign management. An ex post analysis revealed that this led to both (1) a higher success rate and (2) a higher additional return of the campaign.

Table 3. Results of the Evaluation Step regarding the Evaluation Questions E.1 to E.3

A limitation of case-based research is that the results are usually not statistical but analytical generalizations (Yin 2008). According to Lee and Baskerville (2003), theory-informed "rich insight" from single-case analysis is a valid and adequate form of generalized knowledge. Because our analysis has drawn upon a rich case and theoretical foundations, it may be assumed that the observed general results will occur in other cases as well.

4.2 Evaluation of the Metric by Means of Publicly Available Data

The goal of this evaluation step is to analyze the feasibility and the strength of the PBCM based on publicly available data. To do so, we examine the following evaluation questions:

E.4 How can the PBCM be instantiated using publicly available data?

E.5 What is the strength of the PBCM (in comparison with existing currency metrics)?

E.6 How is the strength of the PBCM affected if additional (meta)data are not considered?

4.2.1 Description of the Evaluation Setting and the Datasets

Figure 5 illustrates the proposed general setting:

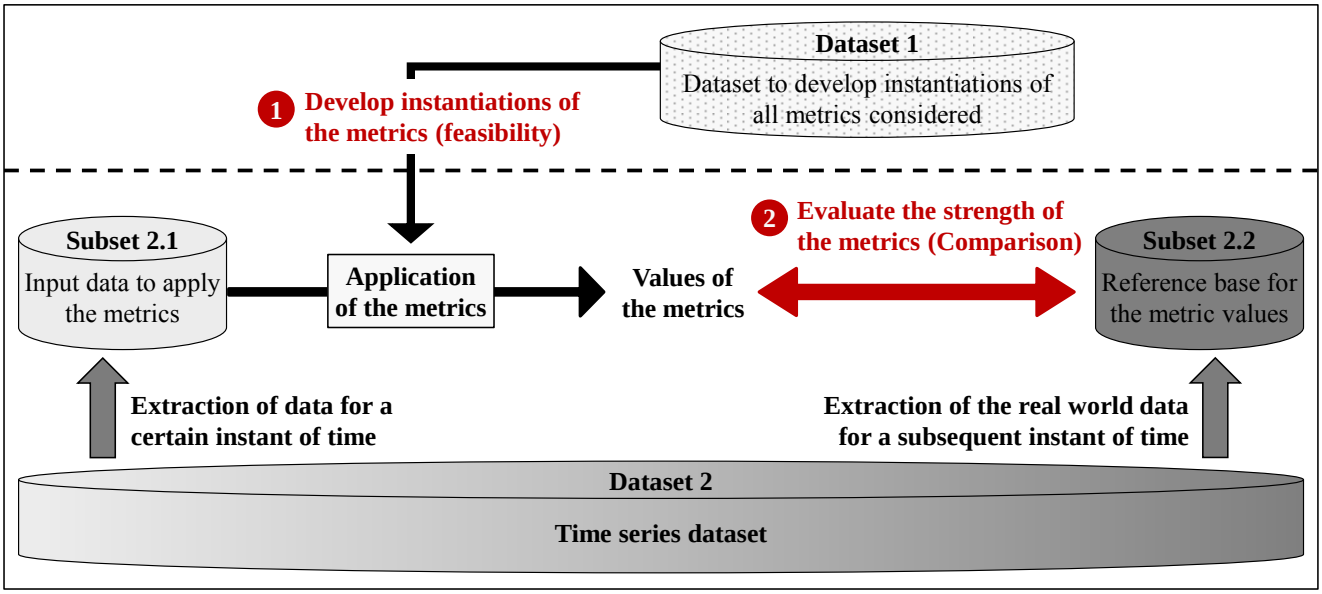

Fig. 5. Evaluation Setting

To address question E.4 and to demonstrate the PBCM's feasibility it is necessary to instantiate the PBCM. Moreover, instantiations of the metrics are needed to be able to evaluate their strength (cf. E.5 and E.6). Hence, a *Dataset 1* is needed which serves as a basis for defining such instantiations. In addition, a further *Dataset 2* is necessary, which contains time series data and constitutes the basis for evaluating the strength of the metrics in the second step (cf. E.5 and E.6). From this real world time series dataset, two subsets have to be extracted. *Subset 2.1* refers to a certain instant of time and shall provide the input data for which the currency has to be assessed for a subsequent instant of time. For example, *Subset 2.1* may contain the data of all persons with the attribute value "student" of the attribute *professional status* regarding the year 2000. Based on these data, the metrics are applied to indicate whether the attribute values "student" of these persons are still current in 2007. *Subset 2.1* shall contain additional data as well. *Subset 2.2* refers to the subsequent instant of time considered and comprises the real world counterparts. In our example, *Subset 2.2* contains the real world data of the attribute *professional status* in 2007 showing whether a person is indeed still a student. Taking *Subset 2.2* as a reference base, it is possible to determine whether an attribute value has changed over time. The strength of the metrics (cf. E.5 and E.6) is evaluated by applying the instantiations of the metrics using the input data provided by *Subset 2.1* and comparing the results to the reference base provided by *Subset 2.2*.

To ensure replicability and repeatability we used publicly available data for *Datasets 1* and *2*. These datasets serve as a basis (1) to instantiate metrics to assess the currency of a person's marital status "single" and (2) to evaluate the strength of the metrics. Considering a person's marital status "single", the currency is affected by a possible marriage. The Statistical Yearbook published by the Federal Statistical Office of Germany (2010) constitutes the most comprehensive statistical reference book in Germany. In addition to others, it includes mar-

riage rates $r_{age,sex}$ (annual number of marriages per 1,000 people) for singles in Germany by age and sex for the year 2008. These data seem appropriate as a basis for instantiating the metrics (*Dataset 1*).

Dataset 2 is required to evaluate the strength of the metrics. The German Institute for Economic Research is the largest economic research institute in Germany and provides time series data collected between 1984 and 2007 in the German Socio-Economic Panel Study (SOEP) for scientific use (cf. German Institute for Economic Research 2008). The SOEP is a wide-ranging representative longitudinal study of private households. Nearly 11,000 households and more than 20,000 persons per year are surveyed. We decided to use them as a basis for evaluating the strength of the metrics (*Dataset 2*). To ensure an adequate sample size and reliable results, we decided to focus on the attribute value "single" of the attribute *marital status*. To illustrate the results of the evaluation step, we chose the year 2000 even though analyses conducted for other years showed similar results. In addition to the marital status, the SOEP dataset also provides data regarding the year of birth and the sex for each person surveyed. These two data attributes denote the main basic demographic, longitudinally tested data within the dataset. They may serve as additional data to apply the metrics and were extracted for *Subset 2.1* as well. In contrast to *Subset 2.1*, *Subset 2.2* serves as the reference base for the values of the metrics. Therefore, it was necessary to extract the marital statuses of all persons included in *Subset 2.1* for a subsequent instant of time. As the data of the year 2007 are the latest available in our SOEP dataset, we use this year's data in the following (analyses conducted for 2001-2006 led to similar results; see Appendix D).

4.2.2 Instantiation of the Metric and Evaluation of its Strength

To demonstrate its feasibility, we instantiate the PBCM to assess the currency of the marital status "single" making it possible to consider a person's *sex* and *age* as additional data W_1 and W_2. We defined the metric as follows. Based on the marriage rates $r_{age,sex}$ provided by the Federal Statistical Office of Germany (2010) (cf. *Dataset 1*), we calculated the probabilities that the marital status "single" of a female or male person ($w_1 \in \{\text{female, male}\}$) with a certain age $t \in \mathrm{N}_0$ is still valid. This was done by $P^{\omega}(T \geq t \mid W_1 = w_1) = \prod_{i=1}^{t}\left(1 - r_{i,w_1}\right)$. To account for the fact that marital status "single" was still current when the person was $W_2 = w_2$ years old, the formula to determine the conditional probability that this person is still single at an age of $t \geq w_2$ can be determined as $P^{\omega}(T \geq t \mid W_1 = w_1, T \geq W_2 = w_2) = \frac{P^{\omega}(T \geq t \mid W_1 = w_1)}{P^{\omega}(T \geq w_2 \mid W_1 = w_1)}$. Hence, the PBCM is instantiated as follows:

$$Q^{\omega}_{Curr.}(t, w_1, w_2) := P^{\omega}(T \geq t \mid W_1 = w_1, T \geq W_2 = w_2) = \frac{\prod_{i=1}^{t}\left(1 - r_{i,w_1}\right)}{\prod_{i=1}^{w_2}\left(1 - r_{i,w_1}\right)} = \prod_{i=w_2+1}^{t}\left(1 - r_{i,w_1}\right) \quad (9)$$

The data provided by *Subset 2.1* contain the marital status "single" that was documented for the year 2000 as well as the year of birth and the sex of 2,978 persons. On this basis, a metric for currency shall indicate whether a person's *marital status* is still "single" in the year 2007. To apply the PBCM denoted in Term (9), it is necessary to calculate its input parameter t as t =2007 - t_0. The SOEP data of the survey conducted in 2007 (*Subset 2.2*) serve as a reference base. To be able to compare the values of our metric in terms of probabilities to the real world facts (i.e., whether the persons are documented to be still single in 2007), we proceeded as follows. First, we categorized the 2,978 persons and assigned each of them to one of the intervals [0; 24], [25; 29], [30; 34], [35; 39], [40; 44], [45; 49], [50; 59], [60; 69], and [70; 100] depending on the age t of the attribute value "single" (i.e., person's age). This was done because the age t of the attribute value constitutes the key input parameter of metrics for currency. The intervals are based on the categorization used by the Federal Statistical Office of Germany (to ensure an adequate sample size in each interval, some categories were aggregated). Moreover, as the deviation of the values for each interval and each metric considered is small, these categories seem to be suitable for our evaluation. Second, for each interval we calculated the average of the metric values. Third, for each interval we determined the fraction of the attribute values "single" that – according to the SOEP real world data – were indeed current in 2007. Finally, both figures were compared as depicted in Figure 6. The SOEP dataset covers 796 persons between 25 and 29 years of age documented to be single in 2000. Referring to the SOEP survey of 2007, 644 (≈80.9%) of them were still single in 2007 (i.e., the attribute value "single" is still current). Calculating the average of the metric values for these persons leads to 0.802 (standard deviation of 0.072). Based on the metric values, this means we expect that approximately 638 (=0.802 · 796) of these persons were still single in 2007. According to Figure 6, the metric seems to provide reasonable indications in terms of probabilities for the actual attribute values' currency. Detailed analyses for the intervals [25; 26], [27; 28], and [29; 30] that contain younger people characterized by considerable decline rates reveal that the fractions of the current attribute values (real world) and the averages of the metric values only differ by two percent or less. Only within the intervals [30; 34] and [35; 39] some differences can be observed. Finally, comparing the intervals' average values of the metric to the fractions of current attribute values in 2007 (real world) and using the sample sizes as weights leads to a mean absolute error (MAE) of 0.0784.

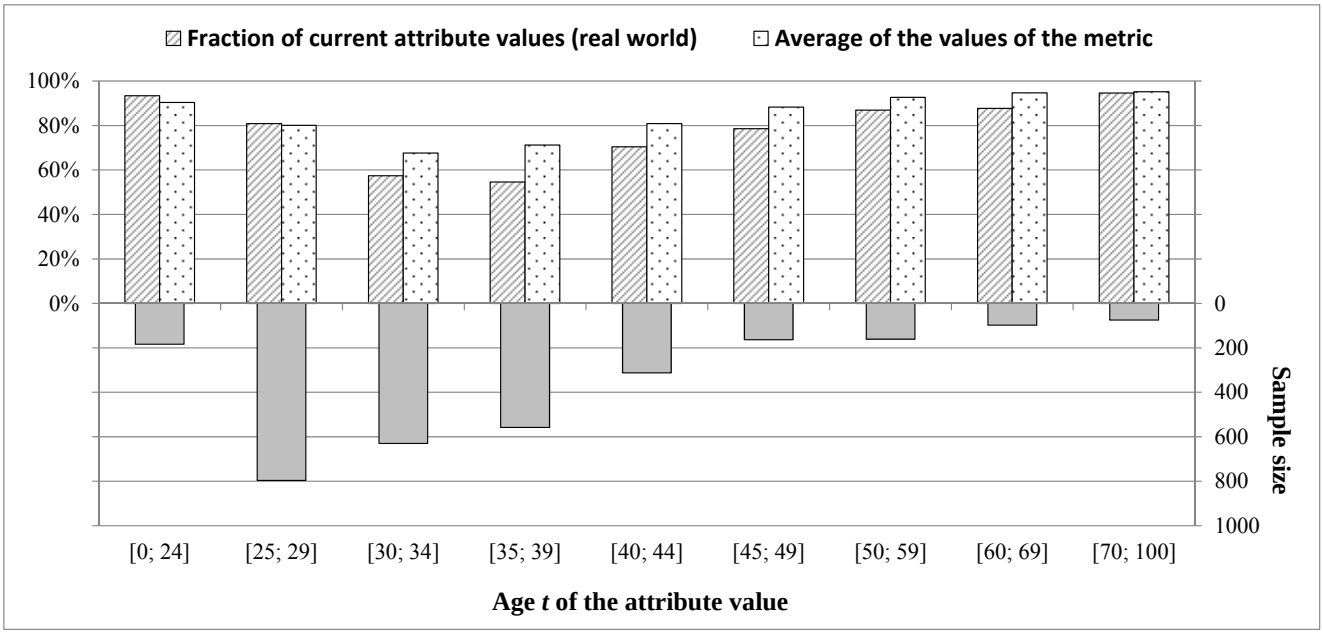

Fig. 6. Fractions of Current Attribute Values and Average Values of the PBCM

Regarding question E.6, how the strength of the metric is affected if additional (meta)data are not known must be analyzed. Thus, we conducted the analysis again, this time neglecting additional data using expected value calculus. If we neglect a person's sex W_1, we have to distinguish two possible events: the person may be female (i.e., W_1=*female*) or male (i.e., W_1=*male*) and the value of the metric $Q^{\omega}_{Curr.}(t, female, w_2)$ or $Q^{\omega}_{Curr.}(t, male, w_2)$. The metric for currency $Q^{\omega}_{Curr.}(t, w_2)$ is defined independent of additional data regarding a person's sex as denoted in Term (10). It constitutes the weighted average of the $Q^{\omega}_{Curr.}(t, female, w_2)$ and $Q^{\omega}_{Curr.}(t, male, w_2)$ using the conditional probabilities of the person being female and male, respectively, as weights g_{female} and g_{male}.

$$Q^{\omega}_{Curr.}(t, w_2) := g_{female} \cdot Q^{\omega}_{Curr.}(t, female, w_2) + g_{male} \cdot Q^{\omega}_{Curr.}(t, male, w_2) \quad (10)$$

Given a 30-year-old person who was single when he or she was 27 years old, the metric value is calculated to $Q^{\omega}_{Curr.}(30, 27) = 0.447 \cdot Q^{\omega}_{Curr.}(30, female, 27) + 0.553 \cdot Q^{\omega}_{Curr.}(30, male, 27) = 0.834$ based on the values for females $Q^{\omega}_{Curr.}(30, female, 27) = 0.808$ and males $Q^{\omega}_{Curr.}(30, male, 27) = 0.855$. Obviously the values of the PBCM are affected if additional data are neglected. Neglecting data regarding the persons' *sex* (W_1), we observe a weighted MAE of the intervals' average values of the metric of 0.095, which is 0.017 (21.6%) higher. The same can be observed if the additional metadata W_2 are neglected.

4.2.3 Comparison with Existing Metrics

To examine question E.5, it is necessary to instantiate the metrics for currency discussed in Section 2.2. In the following, we focus on Ballou et al. (1998), Even and Shankaranarayanan (2007), Heinrich et al. (2007), and

Heinrich and Klier (2011). The strength of the metrics by Hinrichs (2002), Li et al. (2012), and Wechsler and Even (2012) can be discussed based on our analyses as well. Cappiello et al. (2003) and Pernici and Scannapieco (2003) do not aim to assess currency in our sense. Heinrich et al. (2009) do not seek to provide a concrete, mathematically noted metric. The approach by Heinrich and Hristova (2014) focuses on application contexts where (statistical) data are not available and bases on extensive individual expert estimations. Thus, in our evaluation setting an objective comparison is hardly possible (cf. Section 2.2).

To instantiate the metric by Ballou et al. (1998), it is necessary to define the parameter *shelf_life*$\in R^+$ which represents the maximum length of time the attribute value "single" may remain valid. However, specifying an absolute all-time fixed maximum shelf life for this attribute value is obviously not easy. We decided to choose a value of 100 years. Because the sensitivity parameter $s \in R^+$ has to be determined by experts, it was not assigned a fixed value; rather, it was varied to analyze the strength of the metric for different parameterizations and not to limit its strength by the (subjective) determination of a fixed value. The instantiation of the metric was defined as $Q_{\mathrm{B}}(t) := \left\{ \max\left[\left(1 - \frac{t}{100} \right); 0 \right] \right\}^s$ where t denotes the age of the attribute value "single".

We further instantiate both metrics by Even and Shankaranarayanan (2007). The first involves the exponential decline factor *decline_factor*$\in R^+$. The Statistical Office of Germany (2010) provides annual data regarding the number of singles and the number of marriages of singles. On this basis, the parameter was determined to 0.0159. The instantiation of the metric for the attribute value "single" is defined as $Q_{\mathrm{E1}}(t^*) = \exp(-0.0159 \cdot t^*)$ where the age t^* denotes the difference between the years when the currency is assessed and when the attribute value "single" was acquired. To instantiate the second metric, it is necessary to determine the threshold *shelf_life*$\in R^+$ for the attribute value "single". Here, we used a value of 100 years again and defined the instantiation as $Q_{\mathrm{E2}}(t^*) = \begin{cases} 1 - \left(\frac{t^*}{100} \right)^s, & 0 < t^* < 100 \\ 0, & t^* \geq 100 \end{cases}$. For the following analyses, the exponent $s \in R^+$ was not assigned a fixed value but was varied.

The metric by Heinrich et al. (2007) as well as Heinrich and Klier (2011) assumes the shelf life of the attribute value "single" to be exponentially distributed. Its instantiation is equal to the one of the first metric by Even and Shankaranarayanan (2007): $Q_{\mathrm{H}}(t^*) = \exp(-0.0159 \cdot t^*)$.

Because t^*=7 is constant, the first metric by Even and Shankaranarayanan (2007) and the metric by Heinrich et

al. (2007) as well as Heinrich and Klier (2011) result in a fixed value of 0.895 for all 2,978 attribute values "single" in *Subset 2.1*. The second metric by Even and Shankaranarayanan (2007) involves the parameter *s*. Assigning *s* the values 0.05, 0.25, 0.5, and 1.0, the values of the metric are 0.125, 0.486, 0.735, and 0.930. It can easily be shown that the weighted MAE of these existing metrics are much higher than the weighted MAE of 0.0784 of the PBCM. Likewise, it is evident that it is not possible to calculate any (fixed) metric value that results in a weighted MAE smaller than 0.124. Hence, this lower bound for the MAE also holds for the metrics by Hinrichs (2002), Li et al. (2012), and Wechsler and Even (2012).

The metric by Ballou et al. (1998) depends on the age *t* of the attribute value. The results of the analysis of this metric using a sensitivity parameter's value of *s*=1 are depicted in the left diagram of Figure 7.

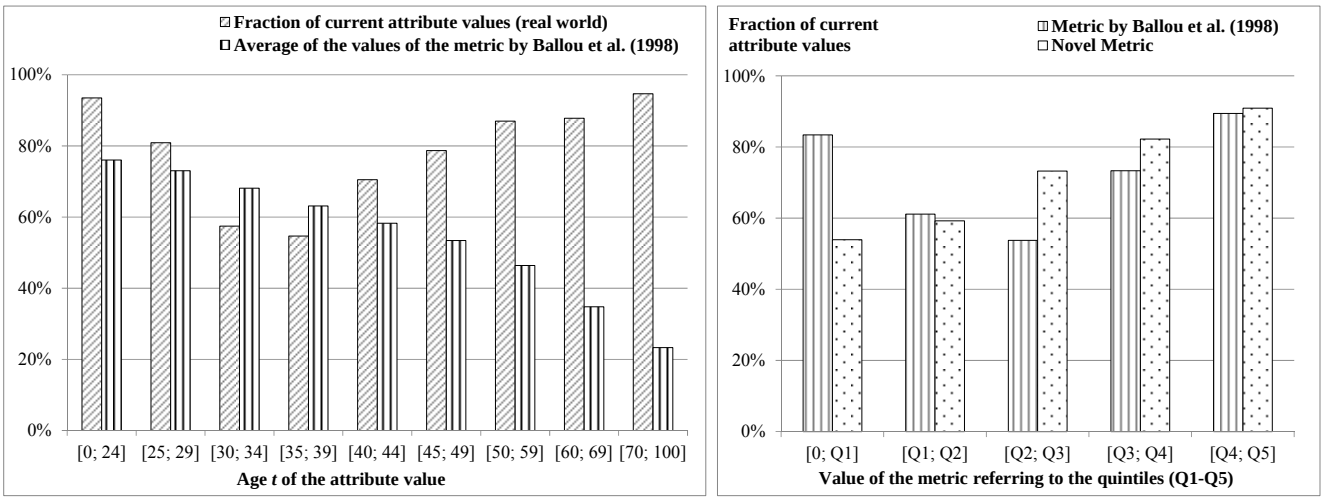

Fig. 7. Fractions of Current Attribute Values and Values of the Metric by Ballou et al. (1998)

The values of this metric strictly decrease depending on the age *t* for any values of the parameters *s* and *shelf_life*, but the fractions of current attribute values do not. That is critical as lower values of the metric shall indicate a lower currency level. To gain deeper insights and to analyze our metric regarding this aspect, we conducted a further analysis. First, we categorized all 2,978 persons depending on their metric value using the respective quintiles (Q1-Q5), once for the metric by Ballou et al. (1998) and once for the PBCM. Then, for each metric and category, we determined the fraction of the attribute values "single" still current in 2007. This way it is possible to elaborate on whether the values of each metric are at least ordinal scaled. This is the case if and only if higher values of the metrics go along with higher fractions of current attribute values (real world). The right diagram of Figure 7 illustrates that higher average values of the PBCM go along with higher fractions of current attribute values. This underlines the fact that the values of the PBCM are at least ordinal scaled (in fact, they are interval scaled). For the metric by Ballou et al. (1998), this does not seem to be the case as the

fractions of current attribute values for smaller values of the metric are higher compared to those observed for higher values of the metric. For example, 83% of the attribute values within the first category "[0; Q1]" are still current although they obtained lower metric values compared to the attribute values within the second category "[Q1; Q2]", while only 61% of these attribute values are still current. Hence, the metric by Ballou et al. (1998) yields in this case higher metric values, even though more attribute values are outdated.

4.2.4 Results of the Evaluation Step

Table 4 summarizes the results of this evaluation step focusing on the feasibility and the strength of the PBCM.

Evaluation Questions	Result
E.4 How can the PBCM be instantiated using publicly available data?	We described how we used publicly available data provided by the Federal Statistical Office of Germany (2010) to instantiate the PBCM taking into account the decline rate and important additional data.
E.5 What is the strength of the PBCM (in comparison with existing currency metrics)?	Applying a publicly available dataset provided by the German Institute for Economic Research (2008), we evaluated the strength of the PBCM. We could demonstrate that the values of the PBCM provide reasonable indications in terms of probabilities for the actual attribute values' currency. Moreover, a comparison with existing metrics revealed advantages of the PBCM regarding the strength and the characteristics of the metric values (e.g., interval scale).
E.6 How is the strength of the PBCM affected if additional (meta)data are not considered?	Analyses based on publicly available data revealed that the strength of the PBCM is affected if additional (meta)data are not considered. When applying the PBCM one has to carefully check which additional data are available or can easily be surveyed and used (considering costs of instantiation) and which additional data are most important with respect to the strength of the metric (cf. Appendix C). Hence, a purposeful assessment of currency is all the more important when applying the PBCM compared to existing metrics.

Table 4. Results of the Evaluation Step regarding the Evaluation Questions E.4 to E.6

However, there are also limitations. First, the overall sample size of 2,978 persons within the SOEP dataset could have been larger to increase the reliability of some analyses. For instance, when comparing the fractions of current attribute values and the average values of the metric depending on the age of the attribute values, the sample size was below 100 for two intervals (cf. Figure 6). To ensure larger sample sizes when analyzing the fractions of current attribute values depending on the metric values, we used the quintiles to determine categories (cf. right diagram of Figure 7). Second, two certain instants of time (2000 and 2007) were chosen to extract input data from the time series data (cf. *Subset 2.1*). Nevertheless, analogous analyses conducted for other years led to similar results, thus supporting our findings (see Appendix D for further analyses).

5. CONCLUSIONS, LIMITATIONS, AND DIRECTIONS FOR FUTURE WORK

Assessing the currency of data in IS is an important issue in science and practice alike. In this paper, we propose a PBCM that is mathematically based on probability theory and makes it possible to assess currency in a

widely automated way. The values of the metric represent probabilities. Hence, they are interval scaled and can be interpreted unambiguously. In fact, the values of the metric can be integrated into expected value calculations in a methodically well-founded manner in order to support decision-making. Compared to existing approaches, it is possible to consider additional data to improve the strength of the metric. Moreover, the PBCM avoids any limiting assumptions regarding particular distribution functions or a fixed maximum shelf life of attribute values. Rather, the metric is applicable for a wide range of attributes and their specific characteristics, such as changing decline rates, and allows for a reproducible configuration (e.g., using statistical methods). The evaluation was conducted in two steps. In a first step, we demonstrated the feasibility, the applicability and especially the practical benefit of the metric in the field of campaign management at a mobile services provider. The adapted customer selection procedure based on the values of the PBCM resulted in considerable economic benefit (higher success rate and additional return). In a second step, the feasibility and especially the strength of the metric were analyzed using publicly available data. It could be shown that the values of the metric provide reasonable indications in terms of probabilities for the attribute values' currency. A comparison with existing metrics revealed advantages regarding the strength and the characteristics of the values.

From a theoretical perspective, it is also worth noting that the characteristics of the PBCM allow for an integration into the theoretical framework of the value of information. The normative concept of the value of information (Carter 1985, Hilton 1981, Lawrence 1999, Marschak et al. 1972, Repo 1989) is defined as the difference between the optimal expected payoff of a decision maker with the respective information versus without using it. The idea behind it is that information is valuable as it can help to reduce uncertainty in the environment of a decision maker and can thus lead to better decisions. Data quality can be described as an "additional layer of uncertainty" (Hilton 1981, p. 62) compared to environmental uncertainty. Against this background, Hilton (1981) applies the theory of comparative informativeness (i.e., Blackwell's Theorem) to illustrate the conditions under which information with higher quality is at least as valuable as information with lower quality. Hilton (1981), however, does not explicitly model the level of data quality. Hence, the PBCM may also be seen as a contribution to be able to explicitly incorporate data quality aspects into the probability-based, normative concept of the value of information. By means of the PBCM, information about the currency of the assessed data can be considered in decision-making and can thus add value in terms of better decisions.

However, there are also limitations that may constitute the starting point for future research. First, our approach is based on the assumption that the probability distribution function of the shelf life of the attribute values con-

sidered can be determined. This assumption is consistent with our synthesis of existing literature (cf. Heinrich et al. 2009) and our practical experiences (cf. Section 4.1). Nevertheless, to substantiate our findings we encourage the definition and the analysis of further instantiations of the metric using different datasets and focusing on different data attributes. Second, in addition to currency, there are further data value-oriented quality dimensions (e.g., completeness) that also need to be taken into account. Indeed, currency is relevant from both scientific and practical perspectives but remains, however, a partial view on the multi-dimensional construct of data quality. Against this background, it is necessary to develop new metrics or enhance existing metrics for other dimensions, which can be integrated into a decision calculus as well. Currently, we are working on a model-based economic approach for planning data quality measures (e.g., data cleansing). This constitutes another step toward an integrated view on different dimensions of data quality. Despite these limitations and directions for future research, we hope that our PBCM will open doors for further research in this exciting area.

REFERENCES

Al-Hakim, L. 2007. Information Quality Factors Affecting Innovation Process. *Int. J. Inf. Qual.* **1**(2) 162-176.

Ballou, D. P., R. Y. Wang, H. L. Pazer, G. K. Tayi. 1998. Modeling Information Manufacturing Systems to Determine Information Product Quality. *Manag. Sci.* **44**(4) 462-484.

Batini, C., C. Cappiello, C. Francalanci, A. Maurino. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.* **41**(3) article 16.

Batini, C., M. Scannapieco. 2006. *Data Quality. Concepts, Methodologies and Techniques*. Springer, Berlin, Germany.

Bureau International des Poids et Mesures. 2006. The International System of Units. http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf.

Cappiello, C., C. Francalanci, B. Pernici. 2003. Time-Related Factors of Data Quality in Multichannel Information Systems. *J. Manag. Inform. Syst.* **20**(3) 71-91.

Carter, M. P. 1985. The valuing of management information Part I: The Bayesian approach. *J. Inform. Sci.* 10(1) 1-9.

Chayka, O., T. Palpanas, P. Bouquet. 2012. *Defining and Measuring Data-Driven Quality Dimension of Staleness (Technical Report # DISI-12-016)*. University of Trento, Trento, Italy.

Cho, J., H. Garcia-Molina. 2003. Effective Page Refresh Policies for Web Crawlers. *ACM Trans. on Database Syst.* **28**(4) 390-426.

English, L. P. 1999. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. John Wiley & Sons, New York, NY.

Even, A., G. Shankaranarayanan. 2007. Utility-Driven Assessment of Data Quality. *Database Adv. Inform.*

Syst. **38**(2) 75-93.

Even, A., G. Shankaranarayanan, P. D. Berger. 2010. Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting. *Decis. Support Syst.* **50**(1) 152-163.

Federal Statistical Office of Germany (2005-2008). *Education and Culture - Statistics of Exams in Higher Education* (in German). Federal Statistical Office of Germany, Wiesbaden, Germany.

Federal Statistical Office of Germany 2010. *Statistical Yearbook 2010 for the Federal Republic of Germany.* Federal Statistical Office of Germany, Wiesbaden, Germany.

Forbes. 2010. *Managing Information in the Enterprise*. Forbes, New York, NY.

Frank, H., S. C. Althoen. 1994. *Statistics: Concepts and Applications.* Cambridge Univ. Press, New York, NY.

Franz, T., C. von Mutius. 2008. *Customer Data Quality - Key to a Successful Customer Dialogue* (in German). *Swiss CRM Forum 2008*, Zürich, Switzerland.

Gelman, I. A. 2010. Setting Priorities for Data Accuracy Improvements in Satisficing Decision-Making Scenarios: A Guiding Theory. *Decis. Support Syst.* **48**(4) 507-520.

German Institute for Economic Research (DIW Berlin). 2008. *The German Socio-Economic Panel Study*. DIW Berlin. Berlin, Germany.

Heinrich, B., D. Hristova. 2014. A Fuzzy Metric for Currency in the Context of Big Data. *Proc. 22nd European Conference on Information Systems*, Tel Aviv, Israel.

Heinrich, B., M. Kaiser, M. Klier. 2007. How to Measure Data Quality? – a Metric Based Approach. *Proc. 28th Intern. Conf. Inform. Systems*, Montreal, Canada, paper 108.

Heinrich, B., M. Kaiser, M. Klier. 2009. A Procedure to Develop Metrics for Currency and its Application in CRM. *ACM Journal of Data and Information Quality* **1**(1) 5:1-5:28.

Heinrich, B., M. Klier. 2009. A Novel Data Quality Metric for Timeliness Considering Supplemental Data. *Proc. 17th European Conf. on Information Systems,* Verona, Italy, paper 240.

B. Heinrich, M. Klier, Q. Görz. 2012. A metric-based approach to quantify the currency of data in information systems. *Zeitschrift für Betriebswirtschaft* **82**(11) 1193-1228 (in German).

Heinrich, B., M. Klier. 2011. Assessing Data Currency - a Probabilistic Approach. *J. Inform. Sci.* **37**(1) 86-100.

Heublein, U., R. Schmelzer, D. Sommer. 2008. *The Development of the Study Dropout Rate at German Institutions of Higher Education* (in German). Hochschul-Informations-System GmbH, Hannover, Germany.

Heublein, U., H. Spangenberg, D. Sommer. 2003. *Reasons for Study Dropouts* (in German). Hochschul-Informations-System GmbH, Hannover, Germany.

Hilton, R. W. 1981. The determinants of information value: Synthesizing some general results. *Manag. Sci.* **27**(1) 57-64.

Hinrichs, H. 2002. *Data Quality Management in Data Warehouse Systems* (in German). Dissertation, University of Oldenburg, Oldenburg, Germany.

IBM. 2012. *Analytics: The Real-World Use of Big Data - How Innovative Enterprises Extract Value from Uncertain Data*. IBM Cooperation, New York, NY.

Klein, B. D., T. J. Callahan. 2007. A Comparison of Information Technology Professionals' and Data Consumers' Perceptions of the Importance of the Dimensions of Information Quality. *Int. J. Inf. Qual.* **1**(4).

Lawrence, D. B. 1999. *The Economic Value of Information*. Springer, New York, NY.

Lee, A. S., R. L. Baskerville. 2003. Generalizing Generalizability in Information Systems Research. *Inform. Syst. Res.* **14**(3) 221-243.

Lee, Y. W., D. M. Strong, B. K. Kahn, R. Y. Wang. 2002. AIMQ: a Methodology for Information Quality Assessment. *Inform. & Manag.* **40**(2) 133-146.

Li, F., S. Nastic, S. Dustdar. 2012. Data Quality Observation in Pervasive Environments. *Proc. 15th IEEE Intern. Conf. Comp. Sci. and Eng. (CSE)*, Nicosia, Cyprus, 602-609.

Marschak, J., R. Radner. 1972. *Economic Theory of Teams*. Yale University Press, New Haven, CT.

Nelson, R. R., P. A. Todd, B. H. Wixom. 2005. Antecedents of Information and System Quality: An Empirical Examination within the Context of Data Warehousing. *J. Manag. Inform. Syst.* **21**(4) 199-235.

Ofner, M. H., B. Otto, H. Österle. 2012. Integrating a Data Quality Perspective into Business Process Management. *Bus. Process Manag. J.* **18**(6) 1036-1067.

Orr, K. 1998. Data Quality and Systems Theory. *Comm. ACM* **41**(2) 66-71.

Parssian, A., S. Sarkar, V. S. Jacob. 2004. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. *Manag. Sci.* **50**(7) 967-982.

Pernici, B., M. Scannapieco. 2003. Data Quality in Web Information Systems. *J. Data Sem.* 48-68.

Pipino, L., Y. W. Lee, R. Y. Wang. 2002. Data Quality Assessment. *Comm. ACM* **45**(4) 211-218.

QAS. 2013. *The Data Advantage: How Accuracy Creates Opportunity*. Experian QAS, London, UK.

Redman, T. C. 1996. *Data Quality for the Information Age*. Arctech House, Boston, MA.

Repo, A. J. 1989. The value of information: Approaches in economics, accounting, and management science. *J. American Soc. Inform. Sci.* **40**(2) 68-85.

Schönfeld, A. 2007. *Address Turntable – Rule-Based Data Exchange with Open Source* (in German). *Open Source Meets Business 2007*, Nürnberg, Germany.

Sidi, F., P. H. S. Panahy, L. S. Affendey, M. A. Jabar,H. Ibrahim, A. Mustapha. 2012. Data Quality: A Survey of Data Quality Dimensions. *Proc. IEEE Int. Conf. Inform. Retr. Know. Manag.* Kuala Lumpur, Malaysia.

Wang, R. Y., V. C. Storey, C. P. Firth. 1995. A Framework for Analysis of Data Quality Research. *IEEE Trans.Knowl.Data Eng.* **7**(4) 623-640.

Wang, R. Y., D. M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inform. Syst.* **12**(4) 5-33.

Wechsler, A., A. Even. 2012. Using a Markov-Chain Model for Assessing Accuracy Degradation and Developing Data Maintenance Policies. *Proc. 18th Americas Conf. on Information Systems*, Seattle, USA, paper 3.

Xiong, M., S. Han, K. Y. Lam, D. Chen. 2008. Deferrable Scheduling for Maintaining Real-Time Data Freshness: Algorithms, Analysis, and Results. *IEEE Trans. on Comp.* **57**(7) 952-964.

Yin, R. K. 2008. *Case Study Research: Design and Methods*. Sage Publications, Thousand Oaks, CA.

APPENDIX A

Symbol / Notation	Explanation
Problem Context and Basic Model of the Probability-based Currency Metric (PBCM)	
ω	Attribute value in an IS
t_0	Instant of creation of the real world counterpart of attribute value ω
t_0'	Instant of capture of attribute value ω in the IS
t_0	Instant when the real world counterpart of attribute value ω changes
t_0''	Instant of acknowledgement or update of attribute value ω in the IS
t_1	Instant of assessing the currency of attribute value ω
t	Age of attribute value ω ($t=t_1-t_0$)
T	Shelf life of attribute value ω (continuous random variable)
w_i, W_i, Ω_{W_i}	Additional data w_i (with $i=1,\ldots,n$) which are relevant when assessing currency of attribute value ω; realizations of the random variables W_i with sample space Ω_{W_i}
w_j, W_j, Ω_{W_j}	Additional data w_j (with $j=1,\ldots,l$ and $l\leq n$) which are known when assessing currency of attribute value ω; realizations of the random variables W_j with sample space Ω_{W_j}
$F^{\omega}(t\|w_1,\ldots,w_n)$	Cumulative distribution function of the shelf life T of attribute value ω depending on the age t of attribute value ω and the additional data w_i
$P^{\omega}(T\leq t\|W_1=w_1,\ldots,W_n=w_n)$	Conditional probability that the shelf life T of attribute value ω is smaller than or equal to its age t given the additional data w_i
$f^{\omega}(\theta\|w_1,\ldots,w_n)$	Density function of the cumulative distribution function $F^{\omega}(t\|w_1,\ldots,w_n)$
$z^{\omega}(t\|w_1,\ldots,w_n)$	Decline rate of the shelf life T of attribute value ω
$Q^{\omega}_{Curr.}(t,w_1,\ldots,w_l)$	PBCM for attribute value ω with age t considering the known additional data w_j
Extensions of the Basic Model of the PBCM	
ω', $W_{\omega'}$	Attribute value ω' stored previously to attribute value ω; realization of the random variable $W_{\omega'}$
$Q^{\omega}_{Curr.}(t,\omega',w_1,\ldots,w_l)$	Extension of $Q^{\omega}_{Curr.}(t,w_1,\ldots,w_l)$ considering attribute value ω' as additional metadata
t'	Age of attribute value ω when it was acknowledged ($t'=t_0''-t_0$)
$Q^{\omega}_{Curr.}(t,t',w_1,\ldots,w_l)$	Extension of $Q^{\omega}_{Curr.}(t,w_1,\ldots,w_l)$ considering the age t' as further additional data
t_0^*	Instant of data entry of attribute value ω (with $t_0\leq t_0^*\leq t_1$)
t^*	Age of attribute value ω with respect to t_0^*, i.e. storage time ($t^*=t_1-t_0^*$)
T^*	Shelf life of attribute value ω with respect to t_0^*
$F^{*\omega}(t^*\|w_1,\ldots,w_n)$	Cumulative distribution function of the shelf life T^* of attribute value ω with respect to t_0^* depending on the age t^* of attribute value ω and the additional data w_i
$f^{*\omega}(\theta\|w_1,\ldots,w_n)$	Density function of the cumulative distribution function $F^{*\omega}(t^*\|w_1,\ldots,w_n)$
$Q^{*\omega}_{Curr.}(t^*,w_1,\ldots,w_l)$	Extension of $Q^{\omega}_{Curr.}(t,w_1,\ldots,w_l)$ based on the age t^* (instead of t)
Evaluation of the PBCM by Means of a Case Study	
r	Additional return in case the customer accepts the offer (estimated to be 6% of the customer's previous sales volume)
R	Total additional return of the campaign

ω	Attribute value stored in the database (here: "student")
t_0	Instant of creation of the real world counterpart of ω (here: *instant of enrollment*)
t_1	Instant of assessing the currency of attribute value ω (here: start of the summer semester 2009)
w_1, W_1, Ω_{W_1}	Additional data w_1 in terms of the respective value of the attribute *type of university* (W_1); Ω_{W_1} ={"University", "University of applied sciences"}
w_2, W_2, Ω_{W_2}	Additional data w_2 in terms of the respective value of the attribute *field of study* (W_2); Ω_{W_2} ={"Economics and Social Sciences", "Engineering Sciences", "Mathematics and Natural Sciences", "Law", "Agriculture, Forestry and Food Sciences", "Education/ Teaching Sciences", "Linguistic/Philology and Cultural Sciences", "Art", and "Health Sciences"}
$P(Dropout \leq t \mid W_1=w_1, \ldots, W_n=w_n)$	Conditional probability that a customer with attribute value "student" has already dropped his or her studies after t semesters
$P(Graduate \leq t \mid W_1=w_1, \ldots, W_n=w_n)$	Conditional probability that a customer with attribute value "student" has already successfully completed his or her studies after t semesters
Evaluation of the PBCM by Means of Publicly Available Data	
ω	Attribute value stored in the database (here: "single" for the attribute *marital status*)
$r_{age,sex}$	Yearly marriage rate for singles depending on person's age and sex
t_0	Instant of creation of the real world counterpart of ω (here: *year of birth*)
t_1	Instant of assessing the currency of attribute value ω (here: 2007)
w_1, W_1, Ω_{W_1}	Additional data w_1 in terms of the respective value of the attribute *sex* (W_1); Ω_{W_1} ={female, male}
w_2, W_2, Ω_{W_2}	Additional data w_2 in terms of the person's age when the attribute value ω was stored (W_2); here: $w_2=2000-t_0$; $\Omega_{W_2}=N_0$
g_{female}, g_{male}	Weights representing the probabilities for being female and male, respectively
shelf_life	Parameter of the metric by Ballou et al. (1998) which represents the maximum length of time the considered attribute value (here: "single") may remain valid
s	Sensitivity parameter of the metrics by Ballou et al. (1998) and Even and Shankaranarayanan (2007)
decline_factor	Decline factor of the metric by Even and Shankaranarayanan (2007)
$Q_B(t)$	Instantiation of the metric by Ballou et al. (1998) depending on the age t of the attribute value "single"
$Q_{E1}(t^*)$	Instantiation of the first metric by Even and Shankaranarayanan (2007) depending on the storage time t^* of the attribute value "single"
$Q_{E2}(t^*)$	Instantiation of the second metric by Even and Shankaranarayanan (2007) depending on the storage time t^* of the attribute value "single"
$Q_H(t^*)$	Instantiation of the metric by Heinrich et al. (2007) as well as Heinrich and Klier (2011) depending on the storage time t^* of the attribute value "single"
Q1, Q2, Q3, Q4, Q5	Quintiles of the calculated metric values

Table A.1. Overview of the Symbols and the Mathematical Notation used

APPENDIX B

Let $(T, W_1,...,W_n)^T$ be a vector of continuous random variables, R^+ the sample space of T, and Ω_{W_i} (with $i=1,\ldots,n$) the sample space of the random variable W_i. Let further be $f(\theta,w_1,\ldots,w_n)$ and $f(w_1,\ldots,w_n)$ the joined probability density functions of the random variables $T, W_1,...,W_n$ and $W_1,...,W_n$, respectively. Then, the conditional probability density function of T given the realizations $w_1,...,w_n$ of the random variables $W_1,...,W_n$ is defined as $f(\theta|w_1,\ldots,w_n)=f(\theta,w_1,\ldots,w_n)/f(w_1,\ldots,w_n)$. The marginal probability density function $f(\theta,w_1,\ldots,w_l)$ of the random variables $T, W_1,...,W_l$ is given by the term $f(\theta, w_1,..., w_l) = \int_{\Omega_{W_n}} ... \int_{\Omega_{W_{l+1}}} f(\theta, w_1,..., w_n) dw_{l+1}...dw_n$.

Based on these identities and assuming that the quotients are defined, for the definition of $f(\theta|w_1,\ldots,w_l)$ it follows that

$$f(\theta \mid w_1,..., w_l) = \frac{f(\theta, w_1,..., w_l)}{f(w_1,..., w_l)} = \frac{1}{f(w_1,..., w_l)} \int_{\Omega_{Wn}} ... \int_{\Omega_{l+1}} f(\theta, w_1,..., w_n) dw_{l+1}...dw_n$$

$$= \frac{\int_{\Omega_{W_n}} ... \int_{\Omega_{W_{l+1}}} f(\theta, w_1,..., w_n) dw_{l+1}...dw_n}{\int_{\Omega_{W_n}} ... \int_{\Omega_{W_{l+1}}} f(w_1,..., w_n) dw_{l+1}...dw_n}, \forall l < n.$$

.

APPENDIX C

We enhanced our metric for currency to deal with unknown additional data for any number of attribute values ω. However, the strength of the metric is affected if (parts of the) additional data are unknown. According to assumption A.2 the distribution of the shelf life T of attribute value ω is given by $F^{\omega}(t|w_1,...,w_n) := P^{\omega}(T \le t | W_1 = w_1,...,W_n = w_n)$. Thus, this distribution function has to be taken into account when analyzing the effect on the strength of the values of our metric. According to assumption A.2, without any loss of generality, the additional data w_k (with $k=l+1,\ldots,n$) are unknown at the instant of assessing currency t_1. As a consequence, we use expected value calculus to remove the unknown additional data w_k from the density function $f^{\omega}(\theta|w_1,...,w_n)$ (cf. Term (1)). Hence, we define our metric based on the cumulative distribution function $F^{\omega}(t \mid w_1,...,w_l) := \int_0^t f^{\omega}(\theta \mid w_1,...,w_l)d\theta$ (cf. Term (2)). This way, it is possible to determine the values of our metric independent from the specific values of the additional data w_k. However, the actual distribution of the shelf life of the attribute value ω considering the additional data $w_1,...,w_n$ is defined by $F^{\omega}(t|w_1,...,w_n)$. Using $F^{\omega}(t|w_1,...,w_l)$ instead affects the strength of the metric.

As the specific values of the additional data w_k are unknown, it is not possible to determine the effect on the strength in a deterministic way. Nevertheless, considering all possible values of the unknown additional data w_k we can determine the maximum absolute error of the value of the metric that may be caused by the missing additional data with respect to the actual probability $P^{\omega}(T \ge t | W_1 = w_1,...,W_n = w_n)$.[3] Mathematically, this maximum absolute error $AE^{\omega}_{\max}(t, w_1,..., w_l)$ can be calculated as follows:

$$AE^{\omega}_{\max}(t, w_1,..., w_l) := \max_{\substack{w_{l+1} \in \Omega_{w_{l+1}}, \\ \ldots \\ w_n \in \Omega_{w_n}}} \left\{ \left| Q^{\omega}_{Curr.}(t, w_1,..., w_l) - P^{\omega}(T \ge t \mid W_1 = w_1,..., W_n = w_n) \right| \right\}$$

$$= \max_{\substack{w_{l+1} \in \Omega_{w_{l+1}}, \\ \ldots \\ w_n \in \Omega_{w_n}}} \left\{ \left| Q^{\omega}_{Curr.}(t, w_1,..., w_l) - \left(1 - F^{\omega}(t \mid w_1,..., w_l, w_{l+1},..., w_n)\right) \right| \right\}$$

Considering the probabilities of the possible values of the unknown additional data w_k as well, the expected absolute error $AE^{\omega}_{\text{expected}}(t, w_1,..., w_l)$ of the value of the metric due to the missing additional data can be determined. Mathematically, $AE^{\omega}_{\text{expected}}(t, w_1,..., w_l)$ is defined as follows:

[3] According to assumption A.2 w_k is a realization of the random variable W_k with sample space Ω_{W_k}.

$$AE^{\omega}_{\text{expected}}(t, w_1, ..., w_l)] := \int_{\Omega_{W_n}} ... \int_{\Omega_{W_{l+1}}} \left| Q^{\omega}_{Curr.}(t, w_1, ..., w_l) - P^{\omega}(T \geq t \mid W_1 = w_1, ..., W_n = w_n) \right| dw_{l+1} ... dw_n$$

$$= \int_{\Omega_{W_n}} ... \int_{\Omega_{W_{l+1}}} \left| Q^{\omega}_{Curr.}(t, w_1, ..., w_l) - \left(1 - F^{\omega}(t \mid w_1, ..., w_l, w_{l+1}, ..., w_n)\right) \right| dw_{l+1} ... dw_n$$

Using the formulas for $AE^{\omega}_{\max}(t, w_1, ..., w_l)$ and $AE^{\omega}_{\text{expected}}(t, w_1, ..., w_l)$ it is possible to get an indication of the strength of the value of the metric for a specific attribute value ω in the case of unknown additional data (an illustration of this fact is provided when presenting the case study).

In addition, it is interesting and important to note that the error measures $AE^{\omega}_{\max}(t, w_1, ..., w_l)$ and $AE^{\omega}_{\text{expected}}(t, w_1, ..., w_l)$ may not only be used to get an indication of the strength of the value of the metric for a specific attribute value ω. Rather, such considerations regarding the strength of the metric can also constitute a useful basis to decide whether or not it is (economically) reasonable to consider specific data attributes as additional data when designing an instantiation $Q^{\omega}_{Curr.}(t, w_1, ..., w_n)$ of the metric. That means they support to trade off the higher reliability of the values of the metric due to (further) additional data against the higher costs for instantiating and applying the metric. Given a metric to assess the currency of the attribute value "student" in a customer database (cf. the mobile services provider of the case study), for example, such considerations may support the decision that it is not economically reasonable to consider the data regarding a customer's marital status as (further) additional data (besides his or her type of university and his or her field of study).

Appendix D

As the data of the year 2007 are the latest available in our SOEP dataset, in our evaluation we focus on the data of this year as a reference base. However, analyses conducted for the years 2006, 2005, 2004, 2003, 2002, and 2001 led to similar results. The results are summarized in the following tables. For all years analyzed the novel metric provides the best results with respect to the MAE.

Reference base: SOEP data of the year 2007 (i.e. t=2007-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	93.48%	80.90%	57.46%	54.66%	70.51%	78.66%	86.96%	87.76%	94.67%	-
Average of the values of the novel metric	90.50%	80.16%	67.68%	71.27%	80.90%	88.33%	92.75%	94.76%	95.23%	7.84%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	87.18%	85.45%	82.51%	79.45%	76.30%	73.05%	68.06%	58.90%	47.87%	15.61%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	76.01%	73.02%	68.08%	63.13%	58.22%	53.37%	46.37%	34.76%	23.31%	15.44%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	57.77%	53.34%	46.38%	39.87%	33.92%	28.50%	21.58%	12.16%	5.75%	29.55%
Average of the values of the metric by Hinrichs (2002)	89.99%	89.99%	89.99%	89.99%	89.99%	89.99%	89.99%	89.99%	89.99%	19.17%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	89.48%	89.48%	89.48%	89.48%	89.48%	89.48%	89.48%	89.48%	89.48%	18.74%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	73.54%	73.54%	73.54%	73.54%	73.54%	73.54%	73.54%	73.54%	73.54%	12.46%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	93.00%	93.00%	93.00%	93.00%	93.00%	93.00%	93.00%	93.00%	93.00%	21.65%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.51%	99.51%	99.51%	99.51%	99.51%	99.51%	99.51%	99.51%	99.51%	28.02%

Reference base: SOEP data of the year 2006 (i.e. t=2006-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	95.48%	78.52%	59.80%	61.40%	73.55%	84.93%	88.32%	89.69%	97.10%	-
Average of the values of the novel metric	92.00%	81.11%	70.69%	75.46%	84.14%	90.45%	93.98%	95.49%	95.99%	7.47%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	87.48%	85.44%	82.44%	79.53%	76.35%	73.00%	68.08%	59.19%	48.33%	14.43%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	76.53%	73.01%	67.98%	63.25%	58.30%	53.29%	46.39%	35.08%	23.72%	14.17%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	58.58%	53.33%	46.23%	40.03%	34.01%	28.42%	21.59%	12.38%	5.94%	31.73%
Average of the values of the metric by Hinrichs (2002)	91.30%	91.30%	91.30%	91.30%	91.30%	91.30%	91.30%	91.30%	91.30%	17.68%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	90.91%	90.91%	90.91%	90.91%	90.91%	90.91%	90.91%	90.91%	90.91%	17.39%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	75.51%	75.51%	75.51%	75.51%	75.51%	75.51%	75.51%	75.51%	75.51%	10.93%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	94.00%	94.00%	94.00%	94.00%	94.00%	94.00%	94.00%	94.00%	94.00%	19.63%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.64%	99.64%	99.64%	99.64%	99.64%	99.64%	99.64%	99.64%	99.64%	24.79%

Reference base: SOEP data of the year 2005 (i.e. t=2005-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	93.90%	81.00%	64.82%	67.33%	78.80%	88.80%	90.48%	89.90%	96.67%	-
Average of the values of the novel metric	93.33%	82.88%	74.41%	80.05%	87.27%	92.32%	95.05%	96.21%	96.71%	5.78%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	87.77%	85.49%	82.42%	79.52%	76.32%	72.95%	68.15%	59.43%	48.50%	11.50%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	77.04%	73.09%	67.94%	63.24%	58.25%	53.22%	46.49%	35.37%	23.87%	14.58%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	59.36%	53.43%	46.18%	40.01%	33.95%	28.35%	21.69%	12.58%	5.99%	34.52%
Average of the values of the metric by Hinrichs (2002)	92.64%	92.64%	92.64%	92.64%	92.64%	92.64%	92.64%	92.64%	92.64%	14.31%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	92.36%	92.36%	92.36%	92.36%	92.36%	92.36%	92.36%	92.36%	92.36%	14.13%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	77.64%	77.64%	77.64%	77.64%	77.64%	77.64%	77.64%	77.64%	77.64%	9.72%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%	95.00%	16.14%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.75%	99.75%	99.75%	99.75%	99.75%	99.75%	99.75%	99.75%	99.75%	20.82%

Reference base: SOEP data of the year 2004 (i.e. t=2004-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	94.89%	83.38%	69.85%	73.96%	85.05%	91.74%	91.23%	90.38%	97.92%	-
Average of the values of the novel metric	94.71%	85.03%	78.73%	84.52%	90.37%	94.19%	96.12%	96.96%	97.46%	4.63%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	88.07%	85.53%	82.47%	79.49%	76.25%	72.85%	68.16%	59.43%	47.93%	9.54%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	77.57%	73.15%	68.02%	63.20%	58.15%	53.07%	46.50%	35.38%	23.25%	16.44%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	60.18%	53.53%	46.28%	39.96%	33.83%	28.19%	21.70%	12.59%	5.65%	37.33%
Average of the values of the metric by Hinrichs (2002)	94.02%	94.02%	94.02%	94.02%	94.02%	94.02%	94.02%	94.02%	94.02%	11.46%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	93.84%	93.84%	93.84%	93.84%	93.84%	93.84%	93.84%	93.84%	93.84%	11.36%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	8.91%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	96.00%	12.99%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.84%	99.84%	99.84%	99.84%	99.84%	99.84%	99.84%	99.84%	99.84%	16.76%

Reference base: SOEP data of the year 2003 (i.e. t=2003-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	95.52%	84.28%	74.74%	81.74%	88.24%	92.73%	91.75%	91.43%	97.62%	-
Average of the values of the novel metric	95.95%	87.78%	83.51%	88.66%	93.11%	95.80%	97.20%	97.74%	98.07%	4.33%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	88.36%	85.54%	82.51%	79.54%	76.35%	72.82%	68.02%	59.81%	48.02%	8.19%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	78.09%	73.17%	68.09%	63.28%	58.29%	53.04%	46.31%	35.83%	23.29%	18.79%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	61.00%	53.57%	46.39%	40.06%	34.00%	28.15%	21.52%	12.92%	5.62%	39.34%
Average of the values of the metric by Hinrichs (2002)	95.45%	95.45%	95.45%	95.45%	95.45%	95.45%	95.45%	95.45%	95.45%	9.18%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	95.35%	95.35%	95.35%	95.35%	95.35%	95.35%	95.35%	95.35%	95.35%	9.13%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	82.68%	82.68%	82.68%	82.68%	82.68%	82.68%	82.68%	82.68%	82.68%	7.08%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	97.00%	97.00%	97.00%	97.00%	97.00%	97.00%	97.00%	97.00%	97.00%	10.64%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.91%	99.91%	99.91%	99.91%	99.91%	99.91%	99.91%	99.91%	99.91%	13.54%

Reference base: SOEP data of the year 2002 (i.e. t=2002-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	97.02%	87.05%	81.82%	87.17%	95.71%	98.08%	94.79%	94.62%	97.37%	-
Average of the values of the novel metric	97.24%	90.94%	88.77%	92.71%	95.70%	97.30%	98.24%	98.49%	98.70%	3.06%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	88.63%	85.49%	82.53%	79.52%	76.40%	72.93%	67.67%	59.98%	48.62%	8.66%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	78.56%	73.10%	68.12%	63.24%	58.38%	53.20%	45.84%	36.02%	23.84%	22.24%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	61.75%	53.45%	46.43%	40.01%	34.10%	28.32%	21.10%	13.04%	5.87%	42.43%
Average of the values of the metric by Hinrichs (2002)	96.92%	96.92%	96.92%	96.92%	96.92%	96.92%	96.92%	96.92%	96.92%	6.22%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	96.87%	96.87%	96.87%	96.87%	96.87%	96.87%	96.87%	96.87%	96.87%	6.21%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	85.86%	85.86%	85.86%	85.86%	85.86%	85.86%	85.86%	85.86%	85.86%	6.51%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	98.00%	98.00%	98.00%	98.00%	98.00%	98.00%	98.00%	98.00%	98.00%	7.15%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.96%	99.96%	99.96%	99.96%	99.96%	99.96%	99.96%	99.96%	99.96%	9.10%

Reference base: SOEP data of the year 2001 (i.e. t=2001-t_0)	[0; 24]	[25; 29]	[30; 34]	[35; 39]	[40; 44]	[45; 49]	[50; 59]	[60; 69]	[70; 100]	MAE
Fraction of current attribute values (real world)	97.81%	91.04%	90.35%	93.82%	98.59%	98.81%	97.96%	96.47%	100.00%	-
Average of the values of the novel metric	98.58%	94.99%	94.28%	96.47%	98.00%	98.75%	99.16%	99.24%	99.35%	2.16%
Average of the values of the metric by Ballou et al. (1998) for s=0.5	88.93%	85.42%	82.60%	79.56%	76.37%	73.02%	67.57%	60.11%	48.88%	11.77%
Average of the values of the metric by Ballou et al. (1998) for s=1.0	79.11%	72.98%	68.24%	63.30%	58.33%	53.32%	45.71%	36.18%	24.06%	25.21%
Average of the values of the metric by Ballou et al. (1998) for s=2.0	62.62%	53.27%	46.58%	40.09%	34.05%	28.45%	21.00%	13.15%	5.94%	44.98%
Average of the values of the metric by Hinrichs (2002)	98.44%	98.44%	98.44%	98.44%	98.44%	98.44%	98.44%	98.44%	98.44%	3.63%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric a)	98.42%	98.42%	98.42%	98.42%	98.42%	98.42%	98.42%	98.42%	98.42%	3.62%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=0.5	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	90.00%	4.88%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=1.0	99.00%	99.00%	99.00%	99.00%	99.00%	99.00%	99.00%	99.00%	99.00%	4.14%
Average of the values of the metric by Even and Shankaranarayanan (2007) - Metric b) for s=2.0	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	99.99%	5.11%